DOCUMENT RESUME

ED 201 642	TM 810 090
TITLE	Construct Validity in Psychological Measurement: Proceedings of a Colloquium on Theory and Application in Education and Employment (Henry Chauncey Conference Center, Princeton, New Jersey, October 1979).
INSTITUTION	Educational Testing Service, Princeton, N.J.; Office of Personnel Management, Washington, D.C.
PUB DATE	Oct 79
NOTE	144p.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Aptitude Tests; *Educational Practices; *Educational Testing; *Employment Practices; Legal Problems; Models; Personality Assessment; *Psychological Testing; Test Theory; *Test Validity; *Theories

ABSTRACT

The stimulus for this colloquium was the convergence of several significant developments bearing on the construct validation of standardized tests and other assessment methods. Of these developments, some were fundamental to psychology as a science; others reflected socio-political pressures on measurement in education and employment. The ten papers presented at the colloquium were: (1) "Foundations of Test Theory" by Charles L. Hulin and Lloyd G. Humphreys; (2) "The Good Theories Do" by Clark Glymour; (3) "Measurement of Abilities Constructs" by John B. Carroll; (4) "Research Models for Exploring Constructs" by Norman Frederiksen; (5) "Application of a Construct Model in Assessment for Employment" by Richard H. McKillip and Hilda Wing; (6) "The Construct Validity of Aptitude Tests: An Information-Processing Assessment" by Robert J. Sternberg; (7) "Construct Validity and Personality Assessment" by Douglas N. Jackson; (8) "Constructs and Their Vicissitudes in Educational and Psychological Measurement" by Samuel Messick; (9) "Legal Issues in Construct Validity" by Barbara Lerner; and (10) "Construct Validity and Validity Generalization" by John E. Hunter. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

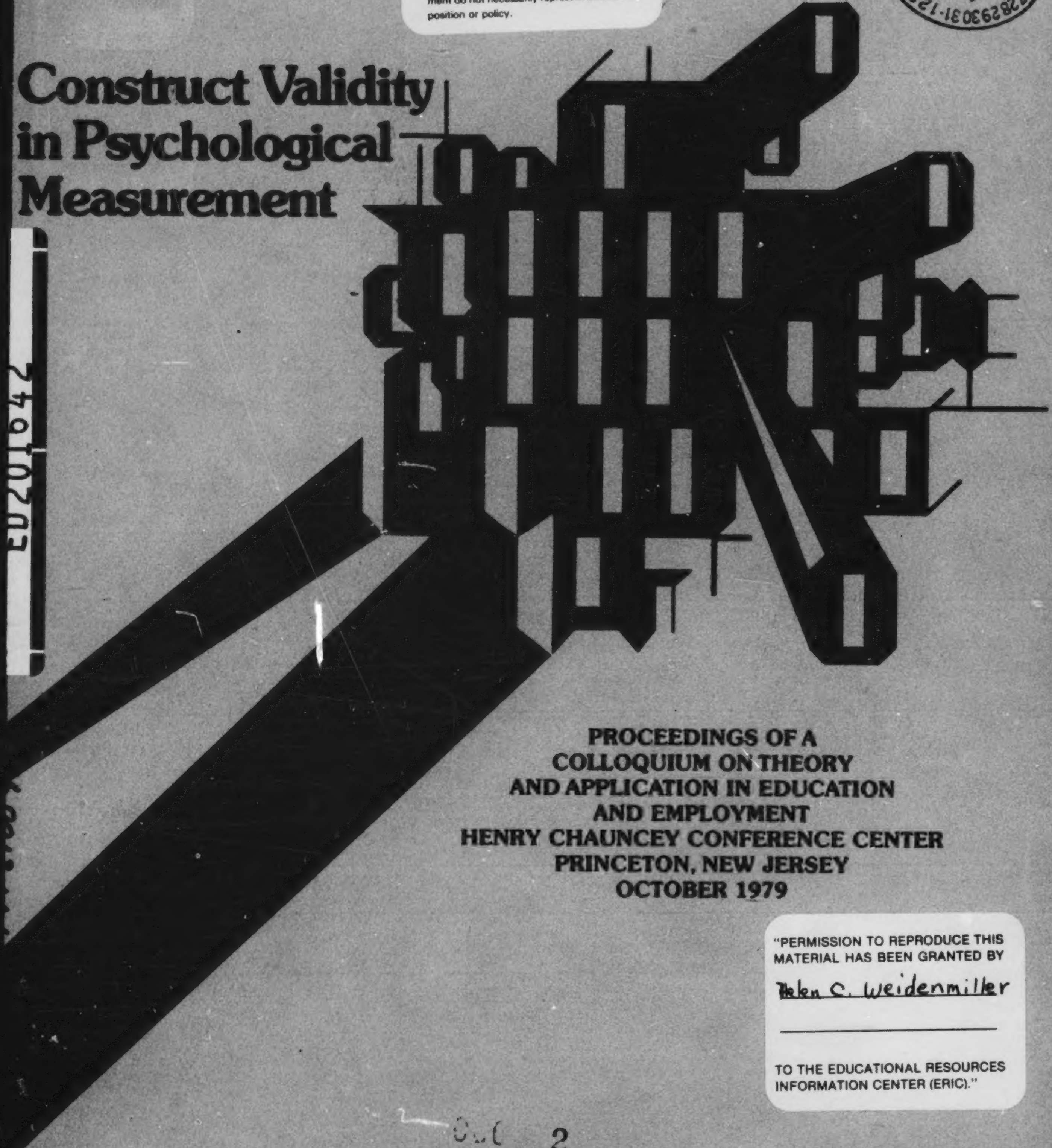
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

RECEIVED
JUL 1980
ERIC/TM

Construct Validity in Psychological Measurement

ED201642



PROCEEDINGS OF A
COLLOQUIUM ON THEORY
AND APPLICATION IN EDUCATION
AND EMPLOYMENT
HENRY CHAUNCEY CONFERENCE CENTER
PRINCETON, NEW JERSEY
OCTOBER 1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Helen C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

2

U.S. OFFICE OF PERSONNEL MANAGEMENT ■ EDUCATIONAL TESTING SERVICE

Funding for this colloquium was provided jointly by
Office of Personnel Management and Educational Testing Service.

Copyright © 1980 by Educational Testing Service. All rights reserved.

Acknowledgements and Editorial Notes

These proceedings comprise the presentations and discussions of the colloquium jointly sponsored by the U.S. Office of Personnel Management and the Educational Testing Service and held at the Henry Chauncey Conference Center, Princeton, New Jersey on 7-8 October 1979.

Planning and organization of this colloquium were primarily the responsibility of William A. Gorham, Richard H. McKillip, and Frank L. Schmidt of the U.S. Office of Personnel Management, and Albert P. Maslow of ETS. Schmidt also contributed very useful and perceptive comments for the Introduction. Ms. Nathale Tullman faithfully recorded and transcribed the entire session, and Ms. Tehirih Foster provided able staff services for the sessions. Maslow, with McKillip's cooperation, and Monica Thatcher's administrative support, managed the editing of these proceedings.

In preparation of these proceedings, transcripts of the presentations and discussion were provided for the authors to use in editing their papers for publication. Some materials which could not be presented during the colloquium because of time limits are included in these proceedings. Generally, the discussions are reported verbatim except for essential editorial clarification.

CONTENTS

E. Belvin Williams
Introduction iv

Arch S. Ramsay
Opening Remarks vii

William A. Gorham
The Setting for this Colloquium 1

Charles L. Hulin and Lloyd G. Humphreys
Foundations of Test Theory 5

Clark Glymour
The Good Theories Do 13

John B. Carroll
Measurement of Abilities Constructs 23

Norman Frederiksen
Research Models for Exploring Constructs 43

Richard H. McKillip and Hilda Wing
Application of a Construct Model in Assessment for Employment 53

Robert J. Sternberg
The Construct Validity of Aptitude Tests:
An Information-Processing Assessment 67

Douglas N. Jackson
Construct Validity and Personality Assessment 79

Samuel Messick
Constructs and Their Vicissitudes in Educational and
Psychological Measurement 93

Barbara Lerner
Legal Issues in Construct Validity 107

John E. Hunter
Construct Validity and Validity Generalization 119

Mary L. Tenopyr, *Chair*
Lee J. Cronbach
John E. Hunter
Robert L. Linn
Panel Discussion on Critical Issues 131

Marvin D. Dunnette
Summary and Integration 139

List of Contributors and Participants 143

INTRODUCTION

It is indeed an exceptional event for 50 behavioral scientists to come together in a collegial atmosphere for a searching examination of a major issue in the measurement of human attributes. It was my special pleasure, therefore, on behalf of William Turnbull, President of ETS, and of Alan Campbell, Director of the U.S. Office of Personnel Management, to welcome the participants and to preside over these sessions.

The stimulus for this colloquium was the convergence of several significant developments bearing on the construct validation of standardized tests and other assessment methods. Of these developments, some are fundamental to psychology as a science; others reflect socio-political pressures on measurement in education and employment. A professional response to these events would seize an unprecedented opportunity to contribute to the conceptual and empirical bases of construct validity and its significance in applied psychological measurement.

The papers presented by the distinguished speakers and panelists and the discussions with their colleagues make a valuable and timely contribution to this objective.

More clearly defined professional standards for construct validation, as Arch Ramsay stresses in his opening remarks, are a clear and pressing need. Gorham points out that vague and inadequate professional standards openly invite lawyers, judges and officials of regulatory bodies to usurp our professional responsibility to define these standards. In the main, such individuals have neither scientific training nor professional interest in psychological measurement. In theory, they should be interested only in enforcement of relevant laws. In practice, however, as Lerner argues, they may have a not-so-hidden extra-legal agenda. Unless psychologists act to establish clear standards in construct validity, we may lose permanently the freedom to do so.

Closely linked to the standards problem is the growing awareness among applied psychologists that to try to base *all* assessment procedures in education and employment on criterion-related and content validity is just not feasible technically, not possible economically, not desirable scientifically. Where group differences in selection rates occur and are not otherwise justified, federal law and regulations require of all employers, public and private, substantial evidence of validity. But recent research by Hunter, Schmidt, and Urry presents convincing arguments that criterion-related studies are much less frequently technically feasible than generally assumed. And, to date, little progress has been reported in the search for alternatives to standardized ability tests. An obvious question then—can construct validity procedures contribute to the solution of this problem? The large scale application of construct validity to Federal employment selection described by McKillip and Wing indicates that it can.

The third development is a resurgence of interest in traits or constructs as a useful and productive theoretical foundation for a science of applied psychology. Human traits, once considered as stable response dispositions, were called into question in the late 1960s and early 70s. However, new evidence for cognitive trait stability has recently emerged. Contrary to previous belief—the belief of many psychologists, but not of the man in the street—relations between traits and real-world performance are quite stable across situations. Hunter argues from recently collected data that the apparent situational specificity of aptitude test validities is due to sampling error and other artifacts. Such findings open the way for the generalization of the validity of constructs across settings and across organizations.

A fourth and particularly significant recent event highlighted by this colloquium is the apparent sharp turn towards the study of cognitive processes being taken by some researchers in the psychology of individual differences. Behaviorist influences in differential psychology, strong in the 60's, are now being questioned as inhibiting, rather than facilitating, research progress. Concurrently, as Glymour points out, an emerging post-positivist philosophy of science supports the role of psychological constructs in cognitive psychology, and the instrumental role of theory in determining when a fact is a fact. Messick examines three existential perspectives on the variables that mediate consistencies in test and non-test behavior and methodological problems in construct validation. The strong research interest in exploring how people process information and solve cognitive problems is exemplified in the contributions of Sternberg, Carroll and Fredericksen. On the question of how best to measure "intelligence," Hulin and Humphreys propose that in test construction, high homogeneity may be incompatible with construct validity, and that test content should be as heterogeneous as possible and cognitively complex. In his research on personality constructs, Jackson also reflects this shift toward a cognitive emphasis in formulating and studying constructs.

Following the presentation and discussion of the prepared papers, a panel, comprised of Mary Tenopyr, Chair, and Lee Cronbach, John Hunter, and Robert Linn, focused on what they perceived as the critical issues emerging in the colloquium. Their comments and interaction with the participants provide a significant contribution to the development of professional standards for construct validation. To close the colloquium, Marvin Dunnette ably summarizes and integrates the wealth of the material, data, and ideas compressed into these two days.

In a post-colloquium reflection, Charles Lawshe observed that the federal uniform testing guidelines focus not so much on what can be inferred about what is being measured as on what can be inferred about other behavior. Therefore, he urges that field practitioners must deal with the linkage between test performance and job performance. He concludes that "the pragmatic necessity for early concensus as to what constitutes generally accepted professional practice imposes an urgent mandate for more examination of, experimentation with, and professional interaction about construct validity strategies which address this type of inference."

Perhaps a useful way for me to close this introduction to the colloquium is to encapsulate Dunnette's main themes: As a part of the scientific enterprise, the study of constructs should be pursued by diverse research strategies. The study of cognitive processes and constructs is particularly promising for the advancement of assessment technology. Both professional judgment and empirical data are critical to application of psychological contructs to practical assessment problems. In such applications, social policy constraints are important, but should not inhibit the "never-ending process" of validation research for advancement of knowledge and practice in psychological measurement.

E. Belvin Williams

ARCH S. RAMSAY OPENING REMARKS

Scotty Campbell has asked me to convey to you his pleasure in sharing the sponsorship of this colloquium of distinguished scholars and his regret that he is not able to be with us today. I feel privileged to substitute for him in making these opening remarks, especially so because the subject to be discussed is one which has important implications for managers both in and out of the federal government. A colloquium on the theoretical underpinning of the method of using psychological constructs to infer the validity of devices to measure mental abilities is of great academic and practical interest. Academically, such discussions hold the promise of leading to a better understanding of the nature of human abilities and of their role in human performance. This understanding is essential to assuring the validity and fairness of selection methods in education and in employment. Practically, they may lead to a well-substantiated construct validity strategy, which will help solve a troublesome problem in personnel measurement. This is the problem of how to match the abilities of large numbers of people to the requirements of many jobs and many academic settings in ways that will be fair to all individual applicants and that will not have an undue adverse impact on any group.

Establishing the constructs that underlie successful performance will allow generalization of validity across jobs. Such generalization would increase the utility of selection devices, effecting not only significant economies in public employment, but also a welcome increase in national productivity.

This is an appropriate time to attempt a more definitive understanding of construct validity. This strategy has been the method of choice for inferring the validity of many examinations ever since World War I. Its principles are explained in most standard texts. It is currently a subject of professional and legal concern: it is a subject of the standards of the Joint Committee on Measurement, and of the principles of the Industrial-Organizational Division of the American Psychological Association. Sections of the federal government's uniform guidelines on employee selection procedures are devoted to the topic; and the U.S. General Accounting Office recently quoted these guidelines in a report on a major federal examination.

Nevertheless, the psychological profession has not yet been able to agree either upon an exact definition of construct validity, or upon the best way of demonstrating it. I hope this colloquium will stimulate thought that will lead to a consensus on how to achieve and demonstrate construct validity. Such a consensus would, of course, represent an important advancement in knowledge. It would also have enormous practical consequences, because better information on which to decide measurement and management issues would be available. Consensus would dispel much confusion in the legal profession, which at present is faced with the task of consolidating many varying opinions on a highly specialized, technical subject. Consensus would also untie many knotty problems in the administration of the uniform guidelines. And it would simplify the task of certification of eligibility for jobs or for higher education, on a nationwide basis.

I am confident that the discussions to be held here will result in practical applications for the improvement of personnel selection. I look forward with interest to reading those results. Thank you all for your willingness to participate in this colloquium, and for your contributions to it. I wish you well in your deliberations.

WILLIAM A. GORHAM

THE SETTING FOR THIS COLLOQUIUM

Today I am speaking to you without an organizational affiliation, since I left the Office of Personnel Management at the end of June. I was, however, working with people at ETS, as Arch and Bel have said, in the early phases of planning this effort that we are in today.

In January of this year, psychologists from the Educational Testing Service and the U. S. Office of Personnel Management agreed upon three objectives for this colloquium. First, the identification of issues relating to the application of construct validity and competency testing; second, the stimulation of professional and scientific activity in clarifying issues and solving problems relating to construct validity so as to develop professional consensus regarding proper applications; and third, the stimulation of research to broaden the body of scientific knowledge and theory as it relates to competency testing.

Almost thirty questions were set down in the areas of the current status of construct validity, issues relating to job-relatedness, validity generalization, relationships with content and criterion-related validity, and minimum and ideal standards. We probably could have set down many more questions. In fact, we did. That alone warns us that if construct validity in competency testing is not to suffer the results of a terminal illness, there will have to be either fewer questions or more answers.

The three objectives were stated in a reasonably appealing manner to assure that a meeting would be attended and that the meeting would be of a high scientific order. But why now in 1979? Why a special colloquium? What's happened?

In football there are only three actions you can take if you have the ball—you can run, pass, or kick. Teams have a game plan, adjusting which of these they do to their own capabilities, those of the other team, and the situation. None of the three methods of moving the football is forbidden by the referees or the spectators because the other team cannot defend effectively against it, but there are sets of rules governing each kind of action, and these are known to all, including the spectators. Despite its violence, the game is orderly.

In the measurement of competency in employment and education, we can play the validation game using criterion-related validity. The rules for this method are reasonably well understood. We can do a pretty good job in content validity even though the rules are still not completely clear. But in construct validity we have been caught without a good rulebook. As players we have unlimited options. The referees and the spectators look at us and it appears as if we are making up this aspect of the game to suit ourselves. While it seems unfair to play that way in a society in which there are limited educational slots and fewer jobs than applicants, it is not unlawful not to have rules, unless the result works to the disadvantage of a legally protected group. Unfortunately, many measures we have used in education and in employment, which are constructs, do work to the disadvantage of one or more of these groups. Thus, without a rulebook, in this case, agreed-upon professional standards with some specificity in adverse impact situations, we face one of three options:

- First: have the rules written for us by others so as to be lawful.
- Second: voluntarily abandon this aspect of the game.
- Third: have this aspect of the game declared illegal by others.

We have come dangerously close to number three, having this option declared illegal and terminated. Some of us perceive construct validity's life to be on a failing respirator. The rescue staff is needed now in the patient's room because employment, psychology, and education use similar measures of ability constructs on occasion. What affects one field will inevitably affect the others. Thus the urgency of this colloquium.

We are now in situation one, having the rules written for us by others so as to be lawful. Those rules are the Uniform Guidelines on Employee Selection Procedures (1978). They are the law as interpreted by three federal agencies. While they purport to reflect the validation requirements of professional standards, where the standards are weak and ambivalent, uncertain, or non-existent, the rules are out of our hands and control and under the aegis of those who control the legal and socio-political structures of this country. This is a distressing situation since psychological inferences regarding a measurement procedure are no longer ours to make exclusively.

Thus, while this colloquium has articulated appealing objectives, the real concern from which these derive is summed up by Barbara Lerner as follows:

> The new guidelines have relented a bit on content validity but still cast a discouraging pall over efforts to establish meaningful forms of construct validity.

That construct validity in employment appears to have a terminal illness is a situation partly of our own creation. It is my hope that this colloquium can lay the foundation for a recovery.

While we cannot worry about yesterday, it is, I believe, useful to understand some of the trilateral history which has brought us to this moment: the professional, legal and socio-political dynamics of the past decade in relationship to construct validity for the measurement of competence. Finally, I will offer some limited opinion on the issue of abandoning this classification of validity altogether.

Some Legal History

Prior to 1972 the federal government was not covered by the Civil Rights Act of 1964. This made life somewhat difficult for federal civil rights enforcement agencies since it appeared that the federal government might be applying different standards (usually described as "weaker") for validity to itself than it was requiring of others. In 1970 two lawsuits were brought against the federal government concerning its testing practices. One ended in the Supreme Court as what we now know as *Washington v. Davis*. As a result of dictum in that decision related explicitly to the other case—*Douglas v. Hampton*—the latter case, regarding a defense of the test based on construct validity, was settled for what amounted to administrative costs. Thus the federal government, despite not having followed the EEOC Guidelines, was successful in these two cases, each involving a different kind of validation strategy.

The federal government was unwilling to apply the impractical 1970 EEOC Guidelines to itself and so developed and applied its own in 1972. But because there is the appearance of disorder, disagreement, and lack of political discipline when the government does not speak with one voice, the existence of two sets of guidelines was intolerable. Consequently the Equal Employment Opportunity Coordinating Council, established by the 1972 Amendment to the Civil Rights Act and consisting of EEOC, Justice, Labor, the Civil Rights Commission and the Civil Service Commission, as one of its first acts ordered its staff to get together and agree upon a common set of guidelines.

I have written elsewhere about the many years of agony that followed. The fortunes of drafts of guidelines followed the results of lawsuits, especially *Washington v. Davis*.

While one of the agreed-upon tenets of any new guidelines was the parity of the validation strategies, EEOC clearly resisted this even after the Supreme Court in *Washington v. Davis* recognized content and construct validity as legitimate coexisters to criterion-related validity. While Justice, Labor and the Civil Service Commission published the Federal Executive Agency Guidelines in 1976, reflecting four years of negotiations, EEOC withdrew at the last minute and republished its 1970 Guidelines the next day. It is clear to me that this demonstration of its impotence in the most important problem before it eventually led to the dissolution of the Coordinating Council in 1979 by a presidential reorganization plan. Uniform guidelines had, in the meantime, been published but only after almost three more years of negotiations.

Why was it so difficult if all parties were merely trying to create a set of guidelines consistent with professional standards?

The Professional Standards

Lawyers and other non-psychologists read the APA Standards, the Division 14 Principles, listened to psychologists, and licked their legal and bureaucratic chops. Never mind what the Court said in *Washington v. Davis*, the disarray in the profession and lack of a set of agreed-upon operations for content and construct validity left these strategies vulnerable and the door open for the government to write its own rules. While three federal agencies came to agreement in 1976 on reasonable requirements for content and construct validity (endorsed by APA) the opening up of discussions in the Carter Administration which was to lead to the Uniform Guidelines forced a great movement toward EEOC's viewpoints. Three facts account for much of the current oppressive situation in regard to construct validity:

1. The APA Standards were not specific to the employment situation.
2. The Division 14 Principles began its one-paragraph discussion of construct validity with the following words: "*The notion* of construct validity . . ." (emphasis added). Construct validity was henceforth branded as a "notion" in employment psychology by those who wanted validation to be as difficult as possible. Paul Sparks later expressed regret over these words publicly, but the damage was done.
3. The APA Committee on Tests and Assessments in a meeting with the Guidelines staff adopted a deferential position to that staff and to the federal government. This Committee suggested that the government should tell it the rules. This Committee's efforts were helpful, however, since the draft of the section on construct validity with which they were working would have been totally fatal if allowed to stand.

I was present when the political principals met to decide aspects of an early draft of the Uniform Guidelines about which there was staff disagreement. In the area of construct validity one of the principals stated: "Oh, we can't have this construct validity. We've got to get rid of it." I doubt that this individual knew anything technical about the issue except that tests of mental traits resulted in large group differences, and that was enough of a reason to "get rid of it." I had been thinking about this colloquium up to that moment. I think that statement precipitated my doing something about it. In an attempt to salvage construct validity, another political principal attached the six-month death noose around the neck of this strategy. That is, employers would be allowed to use construct validity, but if they hadn't eliminated any adverse impact in six months, they would be required to drop it. Fortunately the Guidelines staff was eventually able to turn aside both elimination and the six-month death noose, politically contrived notions, but was so shaken by the potential of these ill-considered decisions that the final Guidelines were seriously affected. And that is the only reason that I bring these stories to you, so you can have some realistic idea of the forces which were and are at work.

Let me continue and be more specific. The staff, having seen the result of presenting impasses to the principals (and I am now referring only to the 1977-1978 period), tacitly agreed that it would settle things at the staff level. Therefore, as is often the case, the strongest voices won out, and they were, at that time, those of civil rights attorneys. These attorneys are not interested in *inferences*, they are interested in *evidence*. They are not interested in mental processes, they are interested in behaviors—things that can be seen, heard, and objectively reported upon in a court of law. Thus, the final version of the Guidelines loads heavily on this litigating point of view.

And just to nail things down further, the Questions and Answers on the Guidelines published in 1979 indicate that ". . . to the extent that there may be differences between particular provisions of the Guidelines and expressions of validation principles found elsewhere the Guidelines will be given precedence by the enforcement agencies. . ." The federal government does not lose litigation gracefully. This language was deliberately contrived to state a position in direct opposition to that of the Supreme Court which, in 1978, affirmed the U. S. loss in *U.S. v. South Carolina*. In that case the lower appeals court had said: "To the extent that the EEOC Guidelines conflict with well-grounded expert opinion and accepted professional standards, they need not be controlling."

The point is that even if more precise professional requirements in the area of construct validity are articulated, we may face some implementation difficulties. Without them we are even more vulnerable.

The Socio-political Forces

We should recognize that the socio-political forces now in control are looking at the results of testing and assuming that, in the case of an employment test, if the results do not produce a representative work force, the tests are both invalid and antithetical to the current socio-political values and goals. There are some misunderstandings and some understandings floating around. For example:

- The experience of those who have some political control is that better tests will probably produce more adverse impact, that they will not solve the problem of employment of minorities. They are probably right for some time to come.
- There is a conflict over whether hiring should be from among the qualified or from the excellent. Long standing societal values are at the crossroads on this one.
- There is a position being forwarded that constructs equal intelligence in the ability area, and that as long as group differences are found, the use of construct measures implies black inferiority. Thus the use of constructs in measurement is tantamount to racism and it has been so articulated. For example, the literature on group differences in tests and in item types was earlier this year labeled as racist by the EEOC. This position may be the most difficult one we have facing us if we are to continue to use and support construct validity in competency testing. But we are here to clarify issues and I will not presume to identify my set.

I would like to make one observation as we proceed. Inferences about tests are made on the basis of professional psychological judgment, except that they aren't any more. It seems as if the inference must now be reasonable to the untrained. This makes life difficult for us when we want to infer construct validity for certain tests and the tests have little, if any, face validity. Legal specialists, politicians, and the public seem to have decreased confidence in psychological inferences which are counter-intuitive. This brings such measures under a great deal of pressure, and the need for an evidentiary basis for our inferences must be more than "reasonable." The pressure shifts away from professional opinion and onto research, from which the professional opinion had better derive anyhow. But if this colloquium ends with the conclusions that "more research is needed," it would be my strong recommendation that it also specify "What are the questions?"

Whatever it is called, there is a profound need to define a practical validation strategy for employers and other users who cannot employ criterion-related or content methods as presently defined. The arguments for why they cannot need not be gone into here. They have been made and, I believe, should be accepted as the point of departure. As a practical matter, employers cannot be impaled on these strategies and their requirements forever or we will find ourselves out of business in terms of being able to help them. If we stick with construct validity, what is needed is an agreement upon the practical minima to satisfy as many of the professional, legal, and socio-political factions and concerns as possible. This is a staggering order and perhaps unachievable.

Might it not be better, then, to rethink the whole organization of validity and perhaps get ourselves out of the mess? Such a proposal is now being given serious thought as many of you know. I'm not going to go into the technical merits of the proposal, but consideration should be given to what the consequences of such a reorganization might be from all of the obvious viewpoints as well as that of public relations. Here are a few of my observations:

- *Washington v. Davis* legitimatized our present organization of validity strategies. Consider what legitimacy a new organization might have, or how it would be achieved.
- Would a new organization suggest that we are trying to be legally evasive?
- Would a new organization be more easily explainable and rational to non-psychologists?
- Are we moving from a fire, air, and water organization to a periodic table of the elements or are we refining the periodic table? If the former, there is more reason to proceed; if the latter, consider whether now is the time for fussing.
- If there is a new organization of validation strategies new federal guidelines will inevitably be precipitated with new documentation requirements. A lot of past work will have to be redone. With no conservative viewpoint currently represented in the federal political power structure we are likely to face greater, rather than reduced burdens. And by conservative, I mean a viewpoint which would opt for less regulation rather than more.

For science to be in conflict with other forces is of course an old story. It would be more convenient if we could proceed without the worry of these forces, but in this day's real world we must keep legal and socio-political concerns in mind. As Bel has said, the time is here for professional standards to be implemented. The purpose of this colloquium can be significant in that regard. With these thoughts in mind, let us begin.

References

Gorham, W. A. Political, ethical and emotional aspects of federal guidelines on employee selection procedures. In W. C. Hamner & F. L. Schmidt (eds.), *Contemporary problems in personnel*. Chicago: St. Clair Press, 1977.

Lerner, Barbara. Tests and Standards Today: Attacks, Counterattacks, and Responses. In R. T. Lennon (ed.), *Impact of Social Change on Testing*. San Francisco: Jossey-Bass, in preparation.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Justice, U.S. Department of Labor, U.S. Department of the Treasury, Uniform Guidelines on Employee Selection Procedures, *Federal Register*, 1978, 43(166), 38925-38309.

Washington v. Davis. 426 U.S. 229 (1976).

DISCUSSION

Schoenfeldt: You referred to Division 14 Guidelines. The second paragraph says, "In view of the lack of substantial literature extending the concept of construct validity to employment practice, no principles for its use are presented here." I guess we can agree there is no substantial literature. Is there any literature at all that bears on construct validity and employment practice, as such?

Gorham: I think there is such literature. If I were still with the Office of Personnel Management, I would turn to my research director and ask him.

McKillip: Our efforts with regard to the PACE examination focused on searching through the literature and trying to find evidence of situations where a construct model was employed. We ended up primarily with a validity generalization approach. That is, we found instances where tests like this, cognitive ability tests, were applied in jobs like the ones we were applying it to.

Gorham: The answer, Lyle, is, I think the literature probably needs to be organized better. I would like to start back in 1917, at least.

Hunter: It seems to me the better known literature is the negative literature; papers by Barrett especially are widely distributed through the legal profession, mostly attacking the idea of construct validity as being totally absurd.

CHARLES L. HULIN and LLOYD G. HUMPHREYS

FOUNDATIONS OF TEST THEORY*

The construct of intelligence, as most other constructs that are of interest to psychologists, probably was generated to explain a broad set of stimulus-response inconsistencies. Whether these were inconsistencies observed in unstructured situations, were casually observed by psychologists, or were codified in masses of correlation or stimulus-response connections is irrelevant for the moment. (We use inconsistencies here because, as Skinner has noted, if a response always follows the presentation of a stimulus we see no need to explain the connection. Also, at the other extreme, totally random connections do not arouse interest.) When things do not happen that are predictably free to happen some of the time, we become interested. In other words, individuals, free to respond as they wished, did not respond consistently when placed in cognitively demanding situations. Some adapted well; they responded adaptively. Others did not.

As psychologists, we have probably retained our interest in intelligence because, under many circumstances, it is related in predictable ways to socially relevant and important criteria. Clinicians, perhaps more than any other group, recognize that IQ or its deviation score is a very important datum to have about the client.

Let us define intelligence, at least for now, as a result of the processes of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills; it is an abstraction. Such a definition, we hope, stresses the breadth and behavioral roots of our interest in the construct. The definition suggests intelligence tests must be composed of a large number of items that are heterogeneous in content and have only moderate item-item correlations if we are to sample the relevant behaviors adequately. If we impose a criterion of high homogeneity it impairs the construct validity of the intelligence tests and converts it into a measure of a phenotypic trait subordinate in the ability hierarchy.

Unfortunately, in the forty years since Thurstone published "The Primary Mental Abilities," psychometricians and factor analysts have lost sight of the general factor in intelligence. McNemar's APA Presidential Address is an obvious counterexample.

Most are aware of Guilford's contemporary model of the structure of intellect. He has three dimensions or types of items. If these are seen as item types, then we must conclude he has not exhausted the possible facets in ability tests. We could add three levels of a speed facet, two different cognitive sets toward guessing manipulated by instructions, three levels of difficulty, two ways of scoring—right or wrong—and two sense modalities, oral and visual. This generates 8,640 item types, although some cells probably could not be filled, particularly those crossed with the oral modality. These are not primary mental abilities, nor are they equivalent to the chemical elements. There will be order among the correlations among such tests. The order could be approximated given a knowledge of the elements any two tests have in common. If we have sampled unevenly, we will generate group factors. These group factors would tell us more about the habits of test constructors than they would about primary mental abilities or about the construct of intelligence.

Now that I have anticipated what I will say during the next thirty minutes, let me try to fill in some of the details.

We are all familiar with the early history of attempts to measure intelligence. Both in Britain and in America—the names of Galton and Cattell immediately come to mind—the early tests of simple processes were unsuccessful. Spearman in 1904 turned to the intercorrelations of school grades as a more useful source of information about intelligence. By 1905 Binet and Simon had developed a preliminary scale measuring intelligence. In 1906 a more complete test in the form of an age scale was produced. The items would be readily recognizable today as typical intelligence test items.

This test and its successors were composed of items, selected from an original pool, that were intentionally complex cognitively. Understanding of abstractions, manipulation of symbols, knowledge of word meaning, and different forms of problem solving were represented. The item pool also included a variety of content crossed with the operations just described. They had verbal, quantitative, figural, and pictorial questions plus questions about actual objects.

In making a selection from the item pool, two criteria have been primary: a steep curve of percent passing as a function of chronological age and a substantial correlation of item response with total score. The heterogeneity of the item pool is reduced only slightly by the use of these criteria and the well-known form of the distribution of IQ's and mental ages when chronological age is held constant is a matter of happenstance grossly overinterpreted by many psychologists. In a standard test of intelligence such as Stanford-Binet, the wide range of item difficulties for a given age group when tested on an age scale of intelligence, in combination with the moderate level of intercorrelations of the items as a function of their heterogeneity, combine to produce approximate normality.

When the items are intercorrelated, even though the intercorrelations are only moderate in size, there is evidence for multiple factors. A unidimensional set of items should have product-moment intercorrelations that form a simple matrix. Problems of factoring binomial data notwithstanding, the tetrachoric correlations should define a single common factor. The heterogeneity of the items, either in terms of a content analysis or a factor analysis, define several group factors. Thus, the total scores on the test are not unidimensional or factorially pure.

---

*Presented by Charles Hulin

The present family of latent trait theories all assume unidimensionality, and, seemingly, the intelligence test is found wanting. It is merely a hodge-podge of various kinds of items lacking in any rationale. Physical measures of height, for example, are not confounded with variance due to weight and vice-versa. We will come back to the point on latent trait analysis a little later.

This factorially complex test provides one of the most important pieces of information about individuals that a psychologist can obtain. Humphreys' evaluation of evidence concerning the validity of the intelligence test is very positive. We also believe that a different approach to test theory provides ample justification of the intelligence test on theoretical grounds.

We start with the assertion that test theorists have not given sufficient weight to the problems associated with the behavioral measurement of a hypothetical trait by a linear combination of multiple items. Test theory should differ radically from classical measurement theory on these grounds. Intelligence, for example, cannot be measured as such. It is measured in a behavioral context. Each item must have some type of content, require some type of operation, and result in some type of product.

The multiple items introduce in test theory the parameter of homogeneity that is absent from classical measurement theory. Let us explore the ramifications of this addition. The ultimate in homogeneity is the perfect Guttman scale; this is an unattainable goal. Lacking the ultimate, test constructors have striven for a high degree of homogeneity within the limits set by the a priori limits of tests, but this a priori conception of the test is not consistent. Thurstone's limits were wider than Guilford's limits are today.

When, and I stress when, a fourth dimension or facet is recognized in the structure of intellect, the a priori conception will be narrowed again. The narrower the a priori conception, the higher the item intercorrelations become. Basically, the decision that certain items belong together in a test while others do not is judgmental. There are no objective-defining psychometric operations to determine what constitutes a test. Psychometric criteria operate within the bounds set by the a priori conception. It is difficult, therefore, to accept the factors that arise from the intercorrelations of present or foreseeable tests as the equivalent of the chemical elements. Invention is a more appropriate term to apply to a factor in mental tests than is discovery.

It is nonetheless true that tests measuring one element in the Cartesian space defined by Guilford's three dimensions of intellect are more homogeneous than tests measuring two or three elements. A common understanding is that the more homogeneous test is likely to be more nearly unidimensional than the less homogeneous test, and therefore, more appropriate for theoretical purposes. The alternative conception is that tests of Guilford's elements are inextricably complex psychologically. If a measure were available for every cell in Guilford's structure, each test would measure all facets. Each would reflect simultaneously content, product, and operation. The conclusion that such tests are pure can be considered no more than a verbal tour de force. Statistical unidimensionality is not identical with psychological unidimensionality.

Let us clarify this conclusion by contrasting two different analyses of test score variance. In classical measurement theory, test score variance is divided into true and error components. I shall follow Loevinger, (1954) however, with a division into three components. She suggests true score, constant or correlated error, and random error, which are in essence the three that we shall use. We wish to rename two of these and thus emphasize the departure from classical theory.

Random error is completely satisfactory, but in place of *constant error* we substitute *nonrandom noise* or *bias sources of variance*. The term constant error suggests that with care the source of error can be avoided. We suggest, however, that this type of variance is intrinsic to the item. Reasoning, comprehension, and so forth, require content. There are many possible response sets; one is inevitable. Items appear in some format or involve a particular operation. There is always a decision concerning the time allowed. Each examinee has a different background of learning and experience that determines item responses. In sum, there is a host of noise or bias sources of variance that determines an examinee's performance on each item. To eliminate all would be impossible; to attempt to eliminate all seems naive.

For true score we substitute *attribute* which will be used as a more general term than *trait* and one less loaded with surplus meaning. A given attribute can be broad, as intelligence, or narrow, as speed in emitting four letter words beginning with *s*. There is no implication in the use of this term with respect to the heredity-environment issue. There is also no implication concerning how the measure of a particular attribute will behave in a factor analysis.

Bias factors are so intrinsic to the item that there will always be nonzero bias variance in a total score. A technique is needed to make bias variance small in comparison to attribute variance. The only possible method of handling the problem is to bring in as many different bias sources as possible in order to minimize their covariances with each other and thus their joint contribution to total variance. Bias only misleads when it leads to covariances and apparent structure.

In contrast, the goal of high homogeneity for a test leads to the concentration of items in a very narrow area, measuring the same set of components, both bias and attribute, with each item. This is, of course, the procedure which Guilford advocates; he measures simultaneously a specific content, a product, and an operation.

Once it is recognized that a test does not have to be as homogeneous as possible and that, in order to measure the attribute in which one is interested, it is necessary to reject the goal of high homogeneity, a very different approach to test construction is opened up. At the same time, assumptions one makes about intelligence as a construct are brought into question.

If each item measures in some small part the attribute desired, but for most items bias and error variance will outweigh attribute variance, it is necessary to specify only that homogeneity be nonzero. Let the rule of test construction be that to assess attributes, it is necessary to specify nonzero homogeneity. Any set of positively correlated items can be added together to form a test. Nonzero correlations among items have causes. It is legitimate for the test constructor to obtain a measure of the effects of those causes. It is reasonable psychometrically to measure an attribute as general as that of intelligence—or even more general, if you wish, for example, organismic or developmental age. Whether this is reasonable psychologically depends on whether the score is useful practically or theoretically. It does not depend on the degree of homogeneity of the items or how the total score behaves in a factor analysis.

It would be highly desirable that there be no clusters of items within a test of low total homogeneity that have substantially higher levels of intercorrelations than the test as a whole. This is not ordinarily possible. If it were, bias sources of variance would become the functional equivalent of random error. The goal becomes, therefore, to maximize the number of clusters, to balance the size of the clusters, and to spread the bias variance among as many sources as possible.

Let us spell out in detail some of the implications of this point of view for latent trait theory. A very broad test can be essentially unidimensional for an attribute such as general intelligence though at a modest level of homogeneity of the items. If bias factors could be sufficiently dispersed throughout the test, perfect unidimensionality could be attained. Unfortunately, small group factors are inevitable. The larger the number of these, and the more even the contribution of each to total variance, the smaller will be the contribution of the bias factors as a whole to total test variance. A measure of the attribute that most closely approximates unidimensionality will of necessity be very heterogeneous from the content point of view.

A test constructor can redefine the desired attribute. Let us suppose that it is narrower than general intelligence, perhaps abstract reasoning (fluid intelligence). All operations other than reasoning, or perhaps inductive reasoning, will now be excluded, but reasoning will be measured by the widest possible selection of items having different content and products. The level of item intercorrelations will be approximately unidimensional with respect to reasoning ability.

Although it is difficult to see why anyone would wish to do so, it is also possible to define a desired attribute such as figure analogies. The total score on this test would probably approximate unidimensionality much more closely than either of the two preceding, and the level of item intercorrelations will be higher. The total score, however, measures only a small aspect of intelligence and a not very broad aspect of reasoning. Psychologically the total score includes analogies and figure variance along with reasoning variance. This state of affairs cannot be revealed by any statistical manipulation since the three sources of variance enter every item.

Thus latent trait theory requires unidimensionality, or close approximations thereto, but it does not require high homogeneity of the items. If we start with an analysis of what we are about in test development and allow the mathematical theory to reflect the psychological analysis, we can proceed on a more productive course in the future than has been true of the recent past.

If our best efforts can only approximate the necessary assumptions of a prevalent mathematical model and if departures from a perfect fit to the model are misleading rather than merely bothersome, some radical changes are necessary.

What light does or could factor analysis shed on this problem? The interpretation of the factors among items in accordance with the model of test construction here being described emphasizes the differences from traditional procedures and suggests a break with history. First-order factors would typically represent bias and would be discarded as of no importance other than that of a check on the test construction methods. Each first-order factor should make an approximately equal contribution to variance if the total score on the test is to be unbiased with respect to the underlying attribute. The latter, that is, the underlying attribute, the construct we wish to assess, would almost always appear in the second or higher order analysis. Higher order factors tend to be the important ones psychologically whether obtained from intercorrelations of items or of total scores and the hierarchical model is appropriate for both.

There are consequences for test reliability and validity that follow from this approach to test construction. First, it is necessary to distinguish clearly between homogeneity and reliability and to recognize the reliability of the total score on a test is confounded with the homogeneity of the test.

Homogeneity of a test is determined by item intercorrelations. The Kuder-Richardson coefficient is a measure of the homogeneity of the total score of the test that reflects both item intercorrelations and the number of items. For a better measure of homogeneity of total score, correction of the Kuder-Richardson coefficient to a standard number of items would be desirable. Reliability is defined as the correlation between test and retest with a minimum time separation or, when memory of items is important, between parallel forms of the same test. These definitions also apply directly to the reliability of items; the reliability of an item is assessed by retesting the same item or administering a parallel form of the item.

Reliability of the total score on a test could be high even with zero homogeneity of the items if the covariances between parallel items in the separate forms of the test were sufficiently close to the item variances. The reliability of the total score increases, however, as item homogeneity increases, parallel item covariances remaining at a constant level. Furthermore, the influence of the test's homogeneity increases as the number of items increase, since the number of covariances increases at a more rapid rate than the number of variances. Thus the model of test construction here being advocated makes it difficult, though not impossible, to achieve high test reliabilities. The goal of high total score reliability is very likely, as a matter of fact, to bias the selection of the items.

The model being proposed keeps the homogeneity of the total test as low as possible within the limits of the attribute being measured. The test constructor must manipulate the number of items rather than their homogeneity to obtain high reliability.

Tests are administered, however, so that inferences can be drawn about measures other than an immediate retest of the same thing. This is the problem of validity and is the issue that originally led us to assess the attribute of intelligence. Although reliability suffers when bias factors are increased in number but decreased in contribution to total variance, validity is typically increased.

This is simply a manifestation of the well-known bandwidth-fidelity paradox. Note that the bandwidth-fidelity paradox is only a paradox when viewed from the perspective of classical test or measurement theory. It is an expected consequence of the model of test construction being advocated here. It merely illustrates that adherence to goals of high test homogeneity may be incompatible with goals of construct validity—incompatible but not paradoxical.

The items in a test used for predictive purposes should have intercorrelations as small as possible as long as each item is related to the criterion, or, in a theoretical development, to the attribute. The model of test construction that we have described, in other words, is congruent with the requirements of obtaining a maximum multiple correlation between predictors and a criterion.

The British hierarchical model of human abilities, which grew out of factor-analytic theory, is readily adapted to the approach to test theory and to test construction that we have described. The attribute the psychologist wishes to measure can be found at any place in the hierarchy. Even for the narrowest of factors, however, the same principles apply: the items in the test should be as heterogeneous as possible within the limitations of the attribute being measured.

It has also become increasingly clear to us that factors among tests are really not very meaningful nor do they reveal anything beyond the original selection of tests to be factored.

In a recent NATO conference, facet analysis and its implications for test theory were developed. We have tried to summarize the implications of that development in the first few pages of this paper.

The measures of the construct of general intelligence developed using this model would be heterogeneous, cognitively complex, have many behavioral references, and would take advantage of the bandwidth aspect of the bandwidth-fidelity distinction. They would probably be related rather broadly and generally to proficiency criteria that have a cognitive element or emphasis and are socially relevant. Differential validity in the original meaning of the term would be the exception rather than the rule. The current emphasis on cultural bias—differential validity, moderator variables, and so forth—in assessments of the construct of general intelligence or ability for purposes of selection or placement would be modified if not abandoned.

This current emphasis was probably caused by some industrial psychologists who had looked at too many validity coefficients based on too small N's and accidental samples. They were convinced of the labile nature of test validities supposedly sensitive to small changes in job requirements. They in turn convinced government agencies of the sensitivity of predictive validity coefficients to minute changes. Such dysfunctional emphasis might be reversible if measures were developed that are more true to the original meaning of the construct of intelligence and the data that generated our interest.

Reference

Loevinger, J. Effect of distortions of measurement on item selection. *Educ. psychol. measmt.* 1954, 3, 441-448.

DISCUSSION

Hunter: There is one error in the paper and that has to do with how reliability should be measured. The test-retest or parallel forms you recommend be used are wrong. Neither should be used given your assumptions on what you call bias factors, or, if I may use a somewhat older term, what Spearman calls specific factors. If you have specific factors in the test and construct a heterogeneous test, what happens is that there will be a high degree of correlation between those specific factors if you go to test-retest correlations.

One beautiful example of this is the 16-PF. I had a long, extended argument with Raymond Cattell a couple of years ago by correspondence in which I finally convinced him—he was worried that if people use internal consistency coefficients, they will constantly come up with bloated specific factors, and I pointed out bloated factors occur with test-retest reliability estimates.

If you look at the parallel forms correlations in the 16-PF, the reliabilities tend to average something like in the sixties. If you look at the test-retest correlations, they tend to fall in the eighties. Test-retest correlations are wrong because they contain the correlations between what you called the bias factors, all the specific item factors. The Kuder-Richardson formula actually reflects the extent of contribution from a general factor, which he intended to measure.

Hulin: I have never seen a correlation that was wrong—unless it was incorrectly computed. I think it's the interpretation that is wrong. If you follow the position that was being advocated in this paper and distribute the bias factor as widely as possible over all the items, maximizing the sources of the bias and thereby minimizing their contribution to the covariance and thus to the total variance of the test, I am not convinced that the test-retest measure of stability or reliability is wrong. It is a measure of the extent to which we have similar rank orders across a short period of time in performance on the test. Bias factors are inherent and are intrinsic to each item.

Hunter: Let me give you a specific example with the 16-PF. In personality tests it's a lot easier to spread the bias factors, as you would call them, because there are so many specific factors; it's quite easy to get two tests where there is relatively little overlap, as opposed to stratified intelligence tests where you have a figural item and a numerical item and so on. So what happens is that you have corresponding bias factors. I am saying ideally you want to use parallel test reliability where you have different bias factors. This is what happens in the 16-PF, and as a result, when you take the parallel forms correlations, they are much lower than the test-retest correlations. That's because parallel form correlations, when you have different bias factors in the two tests, do, in fact, put the bias factors into the error term where they should be, and the test-retest correlation does not.

Hulin: I don't think you are asking for an answer to that.

Schmidt: It was your point that you could get high test-retest correlations, and at the same time have low KR-20 reliabilities under the circumstances that you described. Are you actually advocating test-retest reliabilities as being the appropriate measure of reliability?

Hulin: Yes, the degenerative example would be a blank consisting of three items—age, sex and race—which might have zero covariance, zero homogeneity and high test-retest reliability.

Schmidt: But you are saying the test-retest is a high homogeneity measure, and this is a counterexample because the blank measuring those three items is not measuring any kind of attribute, what you would call an attribute?

Hulin: Only if we maintain a very strong allegiance to the argument that high homogeneity is necessary for measure of a reasonable attribute. If we abandon that and argue for the general nature of the constructs that we are interested in—because very specific ones aren't very often related to socially relevant or important criteria—then you can construct measures of those general attributes that have very low homogeneity, are valid in the sense of being related to general and important criteria, are valid in the sense that they have sampled widely from the behavioral manifestations that originally generated our definition of the construct, and the failure to achieve high KR-20 estimates of homogeneity only speaks to the homogeneity of the items, not to the usefulness or the validity of the test.

Schmidt: But you do have a theoretical goal of measuring an attribute. That means the correlations among the items can't be zero. You are saying they don't have to be high but they can't be zero. And that would be an example of a case where you had high test-retest, but you were not measuring from a theoretical or empirical sense, so it would show you can have high test-retest and still not have a meaningful attribute at all.

Hulin: You are correct only in the degenerative case of zero item-item covariances. With the rule articulated in the paper that there be nonzero covariances, tests with very low homogeneity and KR-20 estimates of reliability would be produced. Their usefulness would be judged by their validity—construct, content, or criterion-estimates. Meaningfulness of attributes does not reside in the homogeneity of their assessments.

Horn: The model that was put in here resembles an equation with predictors that aren't correlated, and it would be nice if you could find a construct. But the problem that I think we are recognizing here is the assumption of an outside criterion, an assumption that you have something to predict that is heterogeneous, and if you pull together these predictors, that would have those qualities. But that's the rub exactly. I mean, you give this sort of academic accounting of what could be. Nothing of help, as I see it, for defining what that criterion would be that you would attempt to predict. And it does no good to talk about facet theory—or, not for me—as being the way that you can put items together. You can put subject tests together and put specific factors together and put biases together. What is lacking in this prescription, it seems to me, is a recognition of a substantive theoretical base for taking all of that varied data that you recognize as being legitimate and making some order out of it such that you could have a genuine construct, a construct that could be predicted with these correlated predictors.

Hulin: This does require the existence of an outside criterion, but there is evidence that there is such a construct that one might want to predict in selection or placement procedures. Something led us to postulate a construct of something—call it general intelligence or neurosis—that can serve as the criterion. If we define the construct, as the data would suggest they should be, in terms of a very broad set of behavioral responses to complex stimuli, then the theory of test construction being advocated says, "sample broadly, sample very heterogeneously." You would, in fact, reflect the general construct, maximize the number of sources of bias, spread them widely throughout the test items. Don't get seduced by the appearance of any "primary" factors that appear in factor analysis. They probably reflect the habits of the test constructor and nothing else.

It is only when you distribute the biases so that they do not lead to the appearance of structure—which suggests either a hierarchical model or second or third order factors—that you are obtaining measures of the general construct or the general attribute, not influenced by bias covariance.

Horn: That's all fine and good, but we are back to ground zero. That's where we were in 1905. We have made no progress with that prescription. There is no way out.

Hulin: I disagree. I would disagree because we have studiously ignored Sir Godfrey Thomson's prescriptions about intelligence and have interpreted first order factors as though they meant something about mental functions of the testees and not the test constructors.

Horn: That just allows any test constructor to say, "I sampled widely with these and these kinds of things that could represent different aspects of intelligence." It allows anyone to construct a test of intelligence anyway they want by just simply saying "I will sample in accordance with this prescription that you are giving." It doesn't solve any problem.

Hulin: That may be correct, but the counter analogy—perhaps in the industrial psychology area—would be that the "proper" procedures allow anybody to construct any measure of any aptitude they want to as long as they want to saturate their measure with items that are merely rewordings of "How satisfied are you with the colors of the wall of the room?" You can ask the question several times and "discover" a factor of satisfaction with the wall color. If we saturate our test battery with a particular type of item, emphasizing one content or another, one operation or another, anybody can obtain and interpret factors that have less and less to do with the general construct we originally postulated to explain complex behaviors.

Horn: I'm not arguing that. I'm not saying one bad solution should be played off against another. I am saying neither solution is a solution of the construct problem. There is no solution in recognizing that Solution A is not a solution.

Hulin: If our definition of construct includes that the construct is a very general heterogeneous construct that is associated with stimulus-response consistencies or inconsistencies across a variety of situations, then test construction procedures that reflect our original interests in the construct would seem to be preferable to test construction theories or procedures that emphasize small homogeneous, relatively worthless primary areas.

Guion: A long time ago we used to more or less routinely take application blanks, identify items on an application blank that were correlated with some external criterion, set up a weighting system for those items that would result in a total score, and that total score, if it cross-validated, would be an effective predictor of that external criterion. We never pretended that we were measuring anything, and if anybody asked us what it measured, we would say we don't know. We are simply predicting. It sounds as if we are confusing the prediction of attributes to something external.

Hulin: There was no requirement in anything that I said that emphasized that we were oriented toward the prediction of an external criterion, because the development started with a theoretical definition of the construct of intelligence. I hypothesized that interest in intelligence had arisen because people observed consistencies and inconsistencies in the way individuals responded to complex situations.

Lloyd is not interested—nor am I interested—in predicting grades as an ultimate criterion of intelligence in school. But it was the consistency in responding to schoolroom situations that originally generated our construct. We could proceed from that, applying the rules of let the item covariances only be nonzero, spread the biases as widely as possible, and we could apply any number of mathematical models, latent-trait analysis, if we wanted to, to analyze the relations of the items to the underlying trait. And at no point have we gotten seduced by the necessity for predicting job success or job proficiency.

A by-product of the item heterogeneity, the distribution of bias sources, and so forth, is to take advantage of the bandwidth part of the bandwidth-fidelity distinction. We would probably end up with a general test with a cognitive emphasis that predicts proficiency criteria, most of which have a cognitive emphasis. In fact, we could probably do it better than we do now.

Anastasi: I would like to try out an implication I seem to be getting out of this. It seems to me that what this implies is that the construct should be defined in terms of non-task items, that is should be defined in terms of something outside the test, which becomes a criterion, but not the criterion in the traditional sense of specific jobs or specific course of study, but rather behavior based on observation and analysis of significant activities within a cultural setting.

If we start out analyzing behavior that way, we would probably generate, for instance, in our culture, a construct attribute that we call verbal ability, quite apart from any test analyses. This would happen if we look at behavior, significant behavior in our culture, we would come up with that construct.

Hulin: Quite likely. That construct would very likely arise. I have no differences with you on that. We would probably also generate a construct called mechanical experience/information/ability—something like that.

Measures that we generated from such constructs that we hypothesize from our observation would be narrower than the general construct of intelligence that I started the paper with. Lloyd has also argued that the measures we generated for those two constructs—the mechanical and the verbal ability—would probably be less useful than general intelligence for predicting socially relevant criteria.

All of these factors are byproducts of the narrowing of the focus and the narrowing of the a priori conception of what we mean by abilities.

Gulliksen: I can see the point of view that anybody can define any old thing as to trait and so on, and go ahead and get various measures of it, but I think I would like to ask about a different possibility, coaching on ability tests.

One of the things, for example, that's of importance in coaching is, do you improve the person's ability just on the items you coach him on, or have you improved what he can do in general.

Of further interest is, if you go ahead and do something that improves the performance on an initial test score, is that person's ability to get the grades and to do the work in the particular school that he's going to, that would select him on the basis of that score, improved also? In other words, the problem is this: Let's just take a batch of tests that measure or group on one factor, take another batch that measures another factor, and so on. Then you take a group of people and after they have taken the test, you drill them on one of the tests that measure, allegedly, a given factor, and see if you can drill them and coach them on this so when they repeat the test they improve their score. Then, check up and find out the extent to which the scores on other tests that measure the same "factor" or on other measures, such as the non-test measures that were being predicted, are improved.

I tried to start an experiment like that some time ago, and it flopped. That is, the Army intelligence groups supported the preliminary study—just to see if I had some material that was at a suitable level for the recruits at Fort Monmouth—and then cancelled the support as soon as I found out that I had gotten some good preliminary material.

Has a study of that sort been done anywhere?

Hulin: I can give you some indirect evidence on that. Ken Alvares at Bowling Green did a dissertation where his manipulation was not coaching specifically on the test, but was putting people into an environment where probably for the first time in their lives they had to rely on cues that they normally were not responsive to. His subjects were a class who were learning to fly an airplane. For the first time in most people's experiences, they had to maneuver in three dimensions simultaneously; from one orientation at 5,000 feet to another orientation at zero altitude. He saturated his test battery, given before and after training, with measures of spatial visualization, spatial orientation, and spatial relations.

He found after ten weeks and approximately 40 hours in the air significant improvements on the nine measures of spatial ability. He did not find an improvement on three tests of mechanical ability, which he had predicted would not improve, and had put in as a factor for control purposes.

I tried the same thing previous to that in the laboratory and could not find a manipulation powerful enough to overcome twenty years of habits. But when you put subjects in an airplane and tell them "Go up to 5,000 feet and go to a strange airport and land there," for the first time they have to rely heavily on those spatial cues. As you might guess, it's hideously expensive and very long and very complex. But that training and coaching on the general construct improved performance on the nine specific measures of the construct.

Gulliksen: Without specific practice on any of those measures?

Hulin: Without specific practice on any of the measures.

Sternberg: I would like to put in a good word for test bias. Sometimes when a source of bias keeps creeping up again and again, we get angrier and angrier and think of more ways to get rid of it, or introduce different sources. I think sometimes that should be a signal that there is something important here we have to deal with rather than dispense with.

The example I have in mind is a set of biases you get on pencil and paper tests that some people have referred to as test-wiseness, planning a strategy for the test, deciding which items to do when or which to do at all; planning your time, checking the process of your work or monitoring the progress of your work; deciding which items you have time to check on; and so on. Those I guess would be biases. I now think they are probably more important than the content that the test is actually measuring. When we went to tachistoscopic testing, we accidentally got rid of a lot of those biases because the items were presented singly rather than in a group. The result as far as I can see is that the correlations started to go down.

I guess the general point I want to make is that you might want to consider seriously whether you want them to be biases or part of the theory of the kinds of behavior you want to study.

Hulin: The work on adaptive testing or tailored testing, that David Weiss among others is doing, presents items one at a time, but not tachistoscopically, and they have not found a concomitant decrease. It may be that the tachistoscopic presentation introduces a very large source of bias that is not present in other criteria.

Gulliksen: I would like to point out that these results you mentioned, where practice on one set of skills transfers to improvement on another, is something that I would cite as counter to the notion that any tester is free to define any old abilities he wishes, and so on. That is, if improvement—if practice on this also makes improvement on that—then you're not free to not define that as an ability, I would say.

Hulin: I guess it is an example of freedom to define anything one wishes as opposed to defining and measuring something useful.

Wing: I have data that speak to the business about taking tests twice. I haven't analyzed them yet, so I can't answer your question. The test is a multiple abilities battery, and I have data both on practice within the testing session and practice across testing sessions. The practice within is only on one of the abilities; the practice across is on all five. The practice effect between testings on all five is greater than you would predict from the sum of the practices on one test or in one testing session; the changes are positively correlated. I don't know what's going on. There might be things besides sheer practice and, of course, the different abilities of different testees. For example, if you were to take the effect of practice in the session, you would say the average difference in the score would be about half the standard error of measurement, but it's in the magnitude of a full standard error of measurement.

CLARK GLYMOUR

THE GOOD THEORIES DO

My position is faintly ridiculous. Over the years psychologists have paid a great deal of attention to what philosophers have said about how to do science. Some psychologists seem to have thought the philosophers of their day to be in possession of a scientific template which had only to be laid down over matters psychological to produce the outline of a science. I think it was a bad thing then, and would be a bad thing today, for psychologists to look to philosophers for instructions as to how to carry out their enterprise. I am not being modest; quite the contrary. The job of philosophers of science has been to try to answer clearly, accurately and with generality a variety of questions about scientific method and its rationale. The questions are not easy ones, and at any moment the array of prevailing philosophical answers is likely to be at best partially baked. It is hard to give precise, insightful and useful answers to questions as to what characterizes a scientific explanation, or what makes a piece of evidence relevant to some hypotheses but not to others, or what functions theoretical terms have, and it is unwise to suppose that the analyses of any philosopher or school of philosophers ought to be taken as gospel. Not gospel, and not training manual either. The notion of learning how to do science from a study of the philosophy of science seems to me roughly as sensible as the notion of learning how to talk from a study of transformational grammar and phonology. If one wants to learn how to talk the thing to do is to hang about with talkers. If one wants to learn to do science, the thing to do is to hang about with scientists.

My sense of inappropriateness will not, however, keep me from offering a few homilies on constructs in science. Philosophers of science are at least as fractious as the practitioners of any other academic specialty, and there is no point to my giving voice to a cacophony. Much of what I shall say is widely believed among philosophers of science and has been for some time; some of it is idiosyncratic. The reader will be able to tell which parts are which by the footnotes (Note 1).

I will forbear from numbering my homilies, but I will list them. The difference between what is observable and what isn't is not very clear and probably not very clarifiable. A lot of what is reported as observed isn't true. Much of the once common worry about the meaning of theoretical constructs (or as I shall say, hereafter, theoretical terms) is confused, or irrelevant to assessing theories, or both. Worry about what evidence bears on what hypotheses and with what weight isn't irrelevant; it is the essential thing. Good theories do more than save the phenomena. Sometimes a body of propositions obtained with the introduction of theoretical terms is as well or better confirmed than any body of pertinent generalizations one could form without such terms.

The topic of this symposium presupposes that there is a distinction among the notions that are used in psychological theories and perhaps in any theory in any branch of science. Some ideas are ideas of "constructs." Constructs are usually characterized only in a negative way: they are entities or features that are postulated by a theory and which are not observable. All of the worry about constructs and their validity presupposes that there is a distinction between what is observable and what is not observable. It has proved hard to draw any very neat distinction in those terms, and philosophers have grown increasingly pessimistic about the possibility of doing so (Note 2). Some of the difficulties are due to the conflicting demands that are put on any philosophical account of the notion of observability. For example, an explication of the notion of observability ought, it was believed, to agree roughly with common usage, but it should also have the result that what is observable can be detected without reliance on any theory.

One discovers quickly enough that what different people consider observable and report as having observed depends on their training and belief, and so very likely on the theories they have committed themselves to. A layman and a meteorologist may understand one another perfectly well when conversing about cold fronts, but looking at a cloud formation the meteorologist might report observing a cold front where the layman sees only clouds. Examples such as this one have suggested to some writers that what is special about the observable is not that people with normal perceptual apparatus actually do recognize observable things without the aid of theory, but that they could learn to do so. Roughly then, observable features and kinds of things are those which a naive subject can be taught to recognize and discriminate by means of a teaching process that is purely ostensive. Such an account is troubled by vagueness: just how reliable must the recognition and discrimination be? Perceptual discriminations are learned in a special context, and the observer can always be misled or fooled by changes in the context or by imitations. Salt crystals are molecules that can easily be recognized, and if someone learns to recognize molecules from examples of salt crystals he may be very good at just that—at recognizing molecules that happen to be salt-crystals—but of course be completely incapable of recognizing other sorts of molecules. Do molecules then count as an observable kind of thing? One might learn to recognize gorillas pretty well but still be fooled by an actor in a gorilla suit: is a gorilla therefore not an observable kind of thing? One could go on.

My own guess, and that of many others, is that "observable" is a word whose use is extraordinarily context-dependent, that it is applied to circumstances or events or things when they can, in a certain context, be recognized

21

fairly reliably by someone whether or not that recognition involves the tacit use of a lot of theory, and when the recognition involves a perceptual act and meets various causal conditions. It would be stupid to infer from this claim that there is no important distinction among quantities corresponding to the distinction between those quantities that are constructs and those that are not. For one thing there are distinctions in the reliability with which a quantity can be determined. Early 19th century chemists knew how to determine lots of things pretty reliably; they didn't know how to determine the weights of atoms. Psychologists are able to determine scores on intelligence tests reliably; I take it that there is some doubt that general intelligence can be determined with equal reliability.

A more important distinction among quantities, I think, has to do with the dependency of the determination of one quantity on the determination of others. Not only do particular determinations of some quantities depend on the determinations of other quantities, at some stages in the development of a science it may happen that a quantity, a construct, can only be measured by measuring certain other "observable" quantities and applying principles of a theory to these measures. Mid-nineteenth century chemists, again, could only determine atomic weights by measuring combining weights and using simplicity assumptions, or by measuring vapor densities and using Avogadro's hypothesis, or by measuring specific heats and using the law of specific heats. Atomic weight determinations were dependent on measurements of these other sorts of quantities. Very much the same sort of thing is at the root of the distinction in psychology between constructs and observables: whatever the construct may be, it counts as a construct exactly because there is no acceptable way of determining its value save by measuring something else and applying some hypothesis.

Denigrating observability and reconceiving the distinction between constructs and nonconstructs as I have done are scarcely novel acts, and like many philosophical results they may seem rather banal. But they do make possible two important insights. First, that the reports of observations, and observational generalizations drawn therefrom, do not constitute an indubitable foundation for knowledge that it is the business of theories to agree with no matter what. And, second, that much of the worry about the meaning of theoretical terms or of constructs is a misplaced and needless worry.

Philosophers and philosophically inclined psychologists have written a great deal about the problem of the "meaning of theoretical terms" or "hypothetical constructs." The idea was that words and phrases that signify observable events, things and features are in fine shape, but the "meaning" of any other word or phrase must somehow be determined in "observation terms." I think virtually every English-speaking philosopher of science would nowadays object to the formulation of this problem. In the first place, it seems unlikely that there is a clear distinction between those things and circumstances that are observable and those that are not, but even if we are willing to settle for a vague distinction, it seems unlikely that it is neatly paralleled by a distinction in language between observation *terms* and theoretical *terms*. But even allowing that there is such a linguistic distinction, what does it have to do with meaning? The idea behind the problem of theoretical terms, once again, was that for sentences containing theoretical terms to be meaningful, there must exist conditions for the truth of such sentences that can be stated in observation terms alone. This seems like a bad idea. There are many functions of language that are encompassed by that vague word, "meaning." They include the reference or signification of words and phrases, the actual usage of words and sentences, and the inferences speakers make from various sentences. None of these features of meaning require that conditions for the truth of theoretical sentences be specified or specifiable in observation terms. It can be objected that someone does not know the meaning of a sentence if he does not know what would count as deciding its truth. But this is incorrect, at least in any ordinary sense of "know the meaning." I know the meanings of a great many sentences about black holes, but I have no idea what observations would decisively prove their truth. In my case, it may be ignorance; but ignorance is inevitable. In 1810 John Dalton could not have said what observations would count decisively for the atomic theory: that depended on too much later discovery and theory. Certainly it would be aberrant to conclude that Dalton did not know the meanings of his own sentences.

One may have the nagging sense that there is still something special about observation and meaning. If someone frequently says "dog" in the presence of cats, we think he does not understand the word or the sentence, and it is tempting to conclude therefore that observation is the arbiter of meaning. But in these cases the importance of observation is that it is an unlimited source of new occasions for asserting beliefs, and if someone's beliefs about dogs (or black holes or anything else) appear sufficiently anomalous, we often conclude that there is a failure of understanding. Knowing the meaning of a sentence does not necessarily require being able to link the sentence with observation; it means sharing a sufficiently large part of the common disposition regarding the use of the sentence.

The concern in psychology and elsewhere with operational definitions reflects an understandable desire to ensure that experimental results will be reproducible. The aim of science is generalization, and generalization may be impossible if a quantity is measured in different ways in different laboratories, all with conflicting results. But it is one thing to insist that researchers be explicit and clear about how they propose to determine a construct, and why that way is an appropriate one; it is quite another thing to insist that for each distinct method of determination, there must be a distinct theoretical notion; to think, that is, that there can never be two distinct ways of measuring the same quantity. I am afraid that the intent of the author of the notion of operational definitions, Percy Bridgman, was the second thing (Note 3). So used, operational definitions are both an historical absurdity and a guarantee of theoretical paucity. An historical absurdity because I doubt that there is a single fundamental theoretical quantity in the history of modern science that was in practice determined in a single way: not atomic weights, not forces, not the space-time metric, not wavefunctions, not anything. That is because the very possibility of testing theoretical generalizations depends on the ability to determine at least some of the variables they contain in more than one way. The decision to regard each method of determination as defining a unique concept is a decision to avoid theory altogether and to treat theoretical terms as simply abbreviations for descriptions of experimental set-ups. There can only be two results of such a decision; either one kids oneself and talks of operational definitions while not actually using one's theoretical notions this way, or else one settles for theoretical triviality. Clarity is an admirable goal but triviality is not the method of choice for attaining it. The only justifications for the recourse to triviality are that, first, theoretical terms have no meaning without operational definitions or that, second, if theoretical terms do have a meaning, if they are not specified by a unique operational definition, then theoretical generalizations containing them must, on the one hand, be less reliable and less well warranted than reports of observations and, on the other hand, have no virtue save that of implying observational reports. I have said enough about meaning, and I turn to observation and theory and their respective reliabilities and virtues.

Observational reports in science aren't usually absorbed by the scientific community in the form of statements as to the position of a voltmeter needle in a certain laboratory at a certain time, or as to the value of the Wechsler-Bellevue score made by Amos Anonymous on a certain day. Such statements often never see print, or if they do, they see print only in conjunction with some determination of a quantity of interest based upon them, whether it be an estimate of the correlation of two variables in a population, or a combining of two elements, or a value for the angle by which light is deflected when passing near the sun. Now such determinations of "observable" quantities are, in one way or another, the basis for the further determination of theoretical quantities and they are likewise the basis on which we judge the empirical adequacy of proposed theories. Such determinations are usually in error, whether from sampling error or systematic error or erroneous assumptions used in the reduction of the data. Very often the error is significant. This homily has been made almost shockingly vivid to me by a series of historical studies I have carried out over the last two years in collaboration with John Earman of the University of Minnesota. Earman and I have been studying the history of the transition from special to general relativity in the years between 1905 and 1920. There was a great deal of observational work done on the theories in those years and much of it gave values for the quantities of interest that physicists now believe to be quite wrong. Einstein's formula for the dependence of mass on velocity was contradicted by observation only a year after it was published. Einstein predicted a shift in spectral lines in sunlight as compared with terrestrial light sources, and from 1914 until 1919 the best solar spectroscopists in the world repeatedly reported observing no such effect. Einstein's value for the deflection of light by the sun's gravitational field was contradicted by more observations than agreed with it. For example, in 1919 the best determination of the light deflection angle was significantly higher than the predicted value. In psychological parlance: using a one-tailed test, the null hypothesis would be rejected at the .10 significance level. It may or may not amuse you to know that this result was heralded as dramatic confirmation of Einstein's prediction.

There is nothing special about relativity. Every other bit of the history of science I know of shows the same thing, if not so dramatically. Take chemistry again: several experimenters reported results in the 19th century contradicting Dalton's law of definite proportions; considerable evidence developed against Dulong and Petit's law of specific heats, and so on. Whether or not the liability of observation to significant error is part of the standard rhetoric of science, everyone more or less knows about it and it is surely one of the reasons why people are willing to tolerate empirical anomalies for so long (Note 4). Observation not only gives results that contradict the regularities we believe to be true, it also gives regularities that aren't true. There are lots of examples. Prout's law, for instance, which maintains that all atomic weights are integral multiples of the atomic weight of hydrogen, seemed as well or as poorly established in 1820 as were the regularities—the law of definite proportions, the law of combining volumes, the law of specific heats—that we now believe to be genuine laws of nature. More rarely, good observers even report seeing things that aren't there at all. In the late 19th century there were many reported sightings of Vulcan, the hypothetical sub-Mercurial planet. Shortly after the discovery of X-rays, a distinguished French physicist, Blondlot, reported the discovery of a new kind of ray, N-rays he called them, and his finding was confirmed in some other laboratories. The French Academy of Sciences awarded Blondlot a prize for his discovery. There are no N-rays, nor are there any other rays which Blondlot was observing and misdescribing. Blondlot and others were seeing nothing at all.

Among the observations in any science at any time there are bound to be many possible views as to which results are due to real empirical regularities and which to error. One of the most important contributions of a good theory is in providing a guide for sorting wheat from chaff, reality from error. I don't mean to say that our theories determine or create the observations. I do mean to say that our theories do guide us in deciding what apparent empirical regularities to give weight to, in our view of what differences between experiment and prediction are tolerable in making judgments as to what experimental results are likely due simply to error. Different theoretical conceptions are bound to give some important differences in the judgments of the quality of the evidence and its bearing. One needs a good theory.

There are arguments—one argument really—that theories and theoretical constructs must always be gratuitous and at best convenient appendages to observations. The argument has carried a great deal of weight in psychology. One finds it in Skinner's remark that if one has a secure equation linking observed variables, it is pointless to detour through unobserved variables. Another version is given more explicitly in Mackenzie's recent book on *Behaviorism and the Limits of Scientific Method*. Mackenzie writes about agnosticism towards theories that:

> There are almost incontestable logical reasons for this suspension or abjuration of judgment. Any such judgment goes beyond any observations which can provide empirical support for the theory, and to that extent is not even in principle confirmable or disconfirmable by reference to such observations. (p.34)

I don't believe such reasons are either logical or incontestable, and I think the philosophical conclusion they have led to is quite wrong. Observations cannot only confirm theories containing constructs, but sometimes, perhaps fairly often, there is more reason to believe theoretical principles than there is to believe any particular piece of evidence, or any particular empirical regularity that forms part of the empirical grounds for the theory. I think this is reasonably obvious. Suppose it were announced tomorrow that the law of definite proportions had been refuted, that conditions had been discovered in which it systematically failed, and that these results had been confirmed independently in several distinguished laboratories. One might be surprised, but one would not, I think, reasonably conclude that it was all up with the atomic theory. Instead, one would expect that some explanation consistent with atomic theory would develop: perhaps that under various conditions only certain isotopes will enter into a chemical reaction. The reason one would be disinclined to abandon atomic theory in the face of this result is that there is so much *other* evidence for the theory. Of course, if it shows anything, this thought experiment only shows that there are imaginable conditions under which one might reasonably abandon the law of definite proportions but not the atomic theory. But I think reflection will show that we do give more credence to established theories than to any particular piece of evidence for them exactly because a well-established theory is confirmed in diverse ways. That, too, I think, is one of the reasons the scientific community is willing to tolerate empirical anomalies for relatively long periods of time. Even if true, of course, this claim does not establish that there are as good grounds for believing a good theory as there are for believing its *entire* body of observational consequences. But I believe that that can happen as well. Perhaps this last thesis is not so banal; I will try to sketch why I think it is nonetheless true.

There are two kinds of ways in which observations give grounds for believing a hypothesis; by testing the hypothesis, or by being explained by the hypothesis. Often the two go together, but not always. First about tests. Suppose we have a theory containing several hypotheses and one or more constructs, and that various claims about observables can be deduced from the theory. Which observations test which theoretical claims? Sometimes the answer is given by scientific tradition, which is to say by meta-theoretical views that transcend any particular theory. Thus gravitational theorists ever since Newton have known that the motions of the planets constitute tests of any proposed gravitational equation, exactly because they believe that such motions are due to gravity. But of course in many contexts there is no adequate tradition or body of belief sufficient to determine the bearing of evidence with appropriate fineness. That is especially true in new sciences and even in old ones when a new theory undermines prior beliefs about the fundamental mechanisms responsible for familiar phenomena.

In the absence of substantive principles determining the relevance of evidence, how is it to be determined? I think one of the most common and defensible procedures goes as follows: Starting with values of measured quantities, and a definite theory, one uses various consequences of the theory to *compute* values for theoretical, unmeasured variables. From a body of such computations one thus obtains values for all of the variables occurring in a particular hypothesis. The body of computations is carried out in such a way that the very structure of the computations does not itself guarantee that the computed values of theoretical quantities will agree with the hypothesis. So the computed values of the theoretical variables will constitute an instance agreeing with the hypothesis to be tested only if the observed variables from which the computations begin have certain values and not others. Values of the observed variables will confirm a hypothesis if they lead, in such a fashion, to values of the theoretical variables that are in accord with the hypothesis. If they lead to values of the theoretical variables that do not accord with the hypothesis, they disconfirm it (Note 5).

I certainly didn't invent this strategy; it ought to be more or less familiar to everyone. It runs throughout the history of science: Newton used it in his argument for universal gravitation, Cannizzaro used it in arguing for his scheme of atomic weights, Eddington used it in his analysis of the classical tests of general relativity. It is common nowadays in the social sciences. Some of its features deserve special note:

1. The strategy makes the testing of a hypothesis *relative to a theory*. Change the theory and one changes what is tested by the same observations. The obvious question of interest in assessing how well a theory is tested, then, is how well the theory is tested *with respect to itself*.

2. The strategy is not holistic: with respect to a fixed theory, a particular subset of the observed variables will typically provide a test of some hypotheses in the theory but not others.

3. The strategy does not depend on how the theory used is presented. A change in the axiomatization of a theory, for example, will have no effect on which hypotheses are tested by which observations with respect to the theory, provided only that the same variables occur in all presentations of the theory. Again, if the theory consists, say, of a system of linear equations, an alternative but mathematically equivalent set of linear equations will give the same confirmation relations.

4. The strategy generates important methodological demands. Simplicity is one; variety of evidence another. Theories with quantities that cannot be determined from observed variables will thus have hypotheses that cannot be instantiated and confirmed. They are just the quantities that people, according to discipline, dismiss as "redundant" or "physically meaningless" or "nonidentifiable." Since in any particular test, some hypotheses of the theory are used to test another hypothesis of the theory, one wants independent tests whenever possible; that is, one wants to arrange theory and evidence so that not every test of A uses B and vice-versa. That is just a demand for a special kind of variety of evidence.

Comparing theories is a complicated business, with many dimensions. A central one is this: *Other things being equal, if a set of observations provides n repeated confirming tests of each member of a set of hypotheses sufficient to entail all consequences of a theory*, T_1, *and that same set of observations provides at most* $m(< n)$ *confirming tests of each member of any set of hypotheses sufficient to entail all consequences of a theory* T_2, *then on the basis of those observations*, T_1, *is to be preferred to* T_2. One can give a technical analysis of how to count tests using this strategy, (see Appendix) but here I will rely on your intuition. The point I wish to make is that this familiar strategy, and the principle of comparison just stated, has the consequence that a theory with constructs can be as well or better confirmed than is the collection of observational consequences of that theory. If we use the preference principle, we should sometimes give a theory as much or more credence as we give to its observational consequences alone. The reason is that, with respect to the theory, values of distinct sets of observables may test the same body of theoretical hypotheses, whereas if we confine attention to the entire set of observational consequences alone, there may be in this set no sentences permitting us to calculate some observables from others. More typically in the social sciences, where theories often consist of systems of linear equations with undetermined parameters, what happens is that by any reasonable count a good theory is as well tested as is its collection of observational consequences, and the preference principle treats the two indifferently. At least, one may say, such theories do no harm. In fact, they often do some good: they explain.

Explanation is a difficult and much contested subject in the philosophy of science, and I don't propose to offer a synopsis. I certainly don't believe that I, or other philosophers, have as yet any idea of the variety of structural patterns that have explanatory force, or why they do. Two things seem clear, however. First, from Copernicus until now, and doubtless before and after, the fact that a theory explains phenomena has been taken as a reason for giving credence to the theory. Second, explaining is more than saving the phenomena; a theory doesn't explain a phenomenon simply because it logically entails it. I will try to describe briefly one of the things some theories do in explaining the phenomena.

Suppose one has an empirical regularity (or even a well-established "theoretical" regularity). A regularity of positional astronomy, say, or the ideal gas law, or whatever. The regularity gives an equation relating sundry variables, but it appears wholly contingent. There is no apparent reason why pressure, temperature and volume *have* to be related as the gas law says. There is no necessity to the thing, it just is that way. Now suppose that a theory *identifies* the observable quantities with certain constructs; for example, the temperature of a gas is said just to be the mean kinetic energy of its constituent molecules and the pressure of a gas is identified with the average change in momentum of the molecules, and so on. If in the original empirical regularity one substitutes for each observable quantity the construct or combination of constructs with which it is identified, there results a new equation, a theoretical equation to which the empirical regularity is reduced by the theory. Now it can happen that the theoretical equation thus obtained is an instance of a *mathematical truth*, a mathematical identity, and therefore something which could not possibly be false. Mathematical identities, and their instances, are necessary truths; there can be no scientific question as to *why* they are true (there can of course be philosophical questions). When this sort of thing happens, the theory explains the empirical regularity, and it does so by showing that, given what the observables really are, the regularity is necessary and the observables could not possibly be connected in any other way.

Like the testing strategy I described earlier, this pattern of explanation runs throughout the history of science. The kinetic theory's reduction of the gas law provides one example, but there are many others: in Copernican astronomy, classical electrodynamics, general relativity, and no doubt many other subjects beyond my ken. These examples, and others I have used, are rather far from psychology, but I see no reason why the same considerations do not apply there or why theories may not do the same good in psychology. Perhaps then, a final illustration somewhat closer to home.

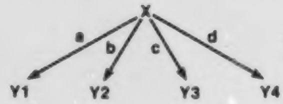
Suppose one is concerned with four measurable variables; let them be, say, scores on subtests of an intelligence test. Denoting the variables by Y_i $(i=1,2,3,4)$ the scores in a sample of the population give six correlation coefficients, ϱ_{ij} $(i,j=1,2,3,4)$. Suppose further that these correlation coefficients are found to approximately satisfy two equations, namely.

i) $\varrho_{12}\,\varrho_{34} - \varrho_{13}\,\varrho_{24} = 0$
ii) $\varrho_{12}\,\varrho_{34} - \varrho_{14}\,\varrho_{23} = 0$

As Spearman's tetrads, I am told, equations such as (i) and (ii) are old hat among psychologists. In consequence, the measured variables will also satisfy other equations which are linearly dependent on i and ii; for example,

iii) $\varrho_{13}\,\varrho_{24} - \varrho_{14}\,\varrho_{23} = 0$

Now one could just leave things as they are, with some interesting but unexplained empirical regularities. The question is, can there be a theory about these phenomena which will do no harm and perhaps some good? Consider the theory which attributes the subtest scores to a single common unobserved factor, plus random error. We can picture the idea by means of a causal diagram:



The letters next to the arrows are to indicate path coefficients. Assume that the variables are standardized, that the dependencies are linear, and that the errors are uncorrelated and have zero mean. Then we obtain a theory given by the following six equations:

1) $\varrho_{12} = ab$ 4) $\varrho_{23} = bc$
2) $\varrho_{13} = ac$ 5) $\varrho_{24} = bd$
3) $\varrho_{14} = ad$ 6) $\varrho_{34} = cd$

Equations 1-6 together imply the observable regularities i, ii and iii. The path coefficients a, b, c and d are theoretical constructs representing the degree of dependency of each of the observed variables upon a common unobserved variable. Now consider the "theory" given just by equations i and ii and their consequences. Any equation of that observational "theory" is tested exactly twice by the determination of a value for each of the six correlation coefficients. Thus, equation i is tested (obviously) by values of ϱ_{12}, ϱ_{34}, ϱ_{13} and ϱ_{24}. But it is also tested by values of ϱ_{34}, ϱ_{13}, ϱ_{24}, ϱ_{14}, and ϱ_{23} in the following way: Use equation ii to calculate a value of ϱ_{12} from values of ϱ_{34}, ϱ_{14} and ϱ_{23}; that is

$$\varrho_{12} = \frac{\varrho_{14}\quad\varrho_{23}}{\varrho_{34}}.$$

The measured values of ϱ_{34}, ϱ_{13}, ϱ_{24} and the computed value of ϱ_{12} now constitute a test of i. There are no other independent tests. It turns out, however, that each of the equations 1-6 is also tested twice by a set of values for the correlation coefficients. Equation 1, for example, can be tested by measuring ϱ_{12}, ϱ_{13}, ϱ_{14}, ϱ_{34}, ϱ_{23}, and ϱ_{24} and computing a and b as follows:

$$a^2 = \frac{(ac)\ (ad)}{(cd)} = \frac{\varrho_{13}\quad\varrho_{14}}{\varrho_{34}} \qquad b^2 = \frac{(bc)\ (bd)}{(cd)} = \frac{\varrho_{23}\quad\varrho_{24}}{\varrho_{34}}$$

Equation 1 will be confirmed in this way provided that

iv) $\varrho_{12}^2\,\varrho_{34}^2 = \varrho_{13}\,\varrho_{14}\,\varrho_{23}\,\varrho_{24}$

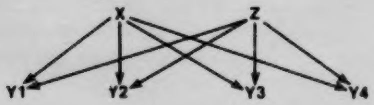
Equation iv is, as is readily seen, just another consequence of i and ii. There is, besides, only one other independent test of equation 1. The case of the other five equations is similar. So far as counting tests is concerned, then, the introduction of the six theoretical equations does no harm.

If we consider explanation, the theory does some good. Equations i and ii, by themselves, seem perfectly contingent. They bear no mark of necessity. But if the correlation between Y_1 and Y_2 just *is* the product of the correlations between Y_1 and X and between Y_2 and X, and analogously for the other observed correlations, then equations i and ii are not accidental; they could not be otherwise. On substituting the theoretical correlations for ϱ_{12}, ϱ_{34}, ϱ_{13} and ϱ_{24} in equation i, for example, that equation becomes

$$(ab)\ (cd) - (ac)\ (bd) = 0,$$

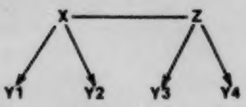
a trivial identity.

It may seem that explaining in this way is a pretty trivial business which any number of alternative theories could accomplish. It is and they can. But it cannot be accomplished without some recourse to theoretical constructs. One may worry that introducing theoretical constructs opens the door to all manner of logically incompatible theories any of which can perform this simple explanatory feat. In some cases a limited indeterminacy may indeed arise, but, if one keeps in mind the simultaneous demand for the best tested theory, the competition is considerably reduced. One cannot, for example, treat the imaginary phenomena of this example nearly as well if two unobserved factors are postulated. A model such as



will turn out to be untestable because the theoretical path coefficients cannot be determined from the observed correlations.

Moreover, it will not imply (i) or (ii). A model such as



turns out to be explanatorily insufficient, since it does not explain both of equations i and ii.

I am sure that the real problems of psychometrics are considerably more complicated than the problems posed by my simple example. I hope there is at least a family resemblance, and that the example suggests the kind of structural relations between theory and evidence that one can hope to have for real psychological theories. I have tried to describe some of the things theories are good for: for assessing observations, for explaining, for providing a well-tested body of at least tentative beliefs. Certainly they are good for other things as well. My only recommendation is to take methodology with a grain of salt. The best of methods is of little help if in some fundamental way one has mistaken what is going on in the world. In science there is no substitute for being right.

Appendix

Consider a theory T consisting of a set of linear equations (and their consequences) with undetermined parameters. Let the evidence be given by a single value of each variable in a subset of the variables occurring in T. The following sequence of definitions will explain how I have counted tests in the text. It is understood that all definitions should, strictly, include the qualifying phrase "with respect to T":

1. A *computation* of a variable or parameter is any equation of T giving that variable or parameter as an (at most finite valued) function of the evidence variables. E. g., a computation of a in the text is

$$a^2 = \frac{\varrho_{13}\quad \varrho_{14}}{\varrho_{34}}$$

2. Let E be an equation and C a set of computations containing one and only one computation of every variable and parameter in E. The *representative of E for C* is the equation obtained by substituting for each variable and parameter in E the function of evidence variables occurring in its computation. For the computation of a given above and

$$b^2 = \frac{\varrho_{23}\quad \varrho_{24}}{\varrho_{34}}$$

and the trivial computation $\varrho_{12} = \varrho_{12}$, the representative of equation I in the text is equation iv.

3. A *test* of E is a set of computations, C, of all variables and parameters occurring in E, such that the representative of E for C is not a mathematical identity, together with a set of values for the evidence variables occurring in the computations.

4. A set, S, of tests of E is *independent* provided no representative of a test in S is a consequence (with respect to real algebra) of the remaining representatives.

5. A set of tests of E is *maximally independent* if it is independent and is not properly contained in any other independent set of tests of E.
6. The number of tests of E is the cardinality of any of the maximally independent sets of tests of E.

Notes

1. An able and interesting defense of views quite different from those I shall state is given in B. Van Fraassen, *The Scientific Image*, Cambridge, Cambridge University Press, forthcoming.
2. See, for example, Maxwell, G. "The Ontological Status of the Theoretical Entities" *Minnesota Studies in the Philosophy of Science*, Vol. III. Feigl, H. and Maxwell, G. eds., Minneapolis, University of Minnesota Press, 1962, and Achinstein, P. "The Problem of Theoretical Terms" *American Philosophical Quarterly*, 2 (1965).
3. See Bridgman, P. *the Logic of Modern Physics*, New York: MacMillan, 1936
4. On the historical tolerance for counter-examples to scientific theories, see Kuhn, T. *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press, 1970.
5. For a more elaborate examination of this strategy of theory testing, see C. Glymour, *Theory and Evidence*, Princeton, Princeton University Press, forthcoming.

DISCUSSION

Horn: I share your view about theory, but as we saw at the introduction today, the problem we are faced with here is not simply a scientific one. It's one of speaking to specifics rather than speaking in general. This kind of background from philosophy of science simply does not exist, the nature of constructs and the support theory simply does not exist. Lawyers do not think that way, as was said before.

Glymour: I think the straightforward remedy is to require all law students to take philosophy courses.

Horn: They know they have the truth in their hands, and science is a bit of an upstart.

Glymour: There is a New York attorney who writes books against evolution. They are quite intelligent and informed. I have talked with biologists about them and the facts are right. But he writes books with titles like "Darwin Retried," and he applies lawyerly canons to Darwinian arguments and always comes up with the result "not proven."

I think that's quite right. It's just a wholly different set of views as to what constitutes good argument, adequate evidence, and so on.

Cronbach: I have an uneasy feeling in terms of the obvious rapid developments in the philosophy of science. But in trying to come down to date, reading as well as I can in the literature, I have been puzzled as to where things have been changing.

I want to say what I think I will change in my paper right now, and I can say I think Paul Meehl is with me. We have been in correspondence and I have been reading his latest. Now I would like to ask you, where else do you think we ought to have a change in the paper? Now, there is a different question.

Glymour: I refuse to answer that question. (laughter)

Cronbach: I think it's a fair question, but you may not have done enough homework on the paper for it to be fair. The first thing that seems to be rather misleading—something we would certainly say differently today—is to put more stress on falsification. We were working in a mode of forward and upward, prior to Kuhn, and today's science is just a better version of yesterday's science.

I think now enough has been said about the way perfectly good theories get abandoned because people like new theories better and we have to go away from this onward and upward style.

The paper ends on the note that construct validation is one among equals. There were in those days four kinds of validation; today there are three. And I think we would both now say the kind of reasoning involved in construct validation is indispensable for all kinds of defensible use of tests, including legal.

I said "the kind"; I am hoping Barbara and I agree so much that she will handle John Horn's comment when the time comes. And the place where most of the hassling about the paper has been, How are constructs defined? At this point Paul says we probably ought to get rid of that word "defined." It causes us nothing but trouble. And where we formerly said a construct is defined by all the sentences that have been written about it, we would now do something more in the vein of saying, Well, now, there are some things that are the core-meaning and there are plenty of other hypotheses about constructs we are ready to abandon.

This is sort of a difference in weight, I think, rather than a reversal of the original claims that were made about what construct validation would consist of.

Now, can you tell me some additional places where we ought to be making changes—not from a legal point of view, but from the point of view of keeping in line with the philosophy of science?

Glymour: No, I can't. What I can do is to indicate to you some sampling of what contemporary philosophers might think about that paper, reading it today, but that's not a recommendation for revision. I think most philosophers of science today think that unless one is extraordinarily careful, talk about meaning and defining things gets people just in a great nest of conceptual confusions. That's only to emphasize more vividly a point you yourself are sensitive to.

Today I think that the philosophers would want to press about the confirmation procedure, the way one does actually confirm sentences containing reference to constructs.

And I think a lot of philosophers would, for example, want you to put the analysis of the baring of evidence in a probabilistic framework, and many philosophers would want it to be a Bayesian framework. I myself wouldn't but in that I'm idiosyncratic; but I promised not to recommend to you what you really should do, only what the consensus of philosophers is.

I think the paper would probably be faulted by philosophical readers today on those two grounds chiefly.

Cronbach: I might add that Paul specifically now is suggesting getting rid of the term "nomological network" and substituting "stochastological" for precisely the point you just made.

JOHN B. CARROLL

MEASUREMENT OF ABILITIES CONSTRUCTS

It gives me much satisfaction to have been invited to say something at this colloquium about what ability tests measure, because I have some strong opinions and beliefs about human abilities, their importance in society, the desirability of having ways of measuring them, and the techniques by which they are or should be measured.

It is absolutely self-evident that, for whatever reasons, there are wide differences in people's abilities, and that those abilities have a lot to do with the way people function in society. Perhaps it is not so self-evident that society has a continuing need for ways of assessing these abilities—the best that we can devise—but I believe a good case can be made for this also. I shall not try to make this case, since that is really not my topic. But certainly, before we can even begin to debate the desirability of having and using ability measurements to guide society's use of people's talents, we need to have a clear understanding of what we mean by ability, what human abilities are or consist of, and how we can measure those abilities. And that is really my topic, for the meaning of ability is centrally involved in any discussion of the "construct validity" of ability tests. What I intend to do is review and discuss both "classical" and newer methods of explicating the construct validity of tests of abilities.

Before doing this, however, it is imperative to step back a bit and consider exactly what we mean by construct validity—or at least, what *I* am going to mean by it in what follows. I do this at the risk of being redundant with, or in disagreement with, the presentation by Professor Glymour and also at the risk of losing my audience with what it may feel is a rather dull philosophical exercise. To make my discussion a little livelier than it might be otherwise, I'm using some graphics and cartoonlike diagrams. Here is the first one:

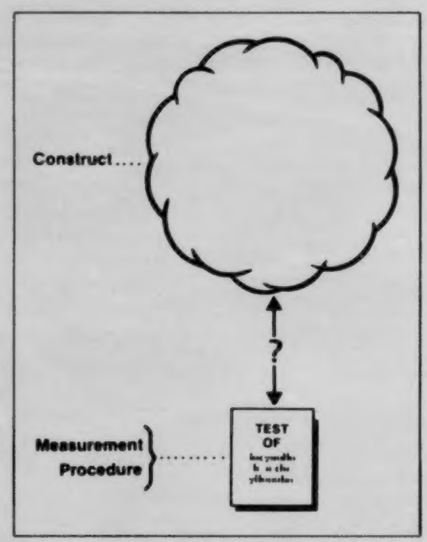


Figure 1.

The concept of construct validity assumes that in any given case, there are at least two elements to be considered: (1) some sort of "construct," and (2) some measurement or assessment procedure. The problem (Figure 1) is to establish a satisfactory connection between the two.

But this is just the simplest case. Relations between constructs and measurements may be much more complicated. In reality, a certain test might have relations with several constructs, as shown here (Figure 2):

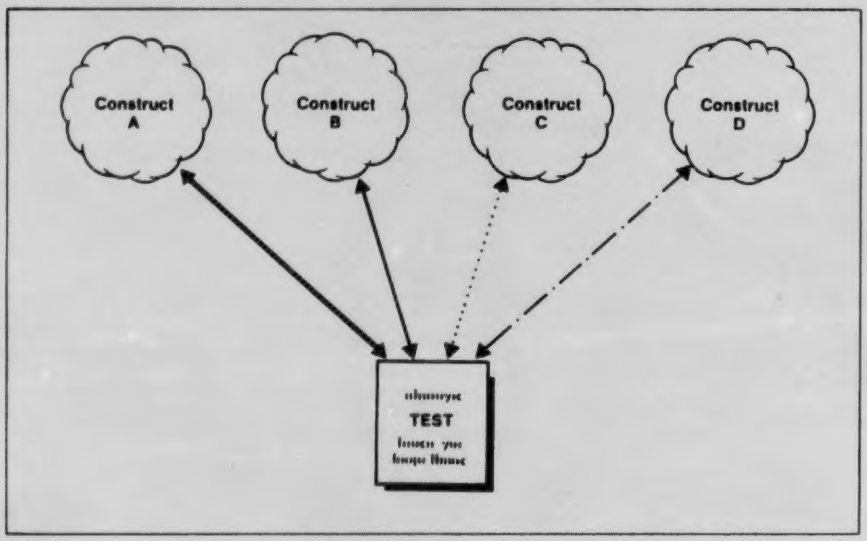


Figure 2.

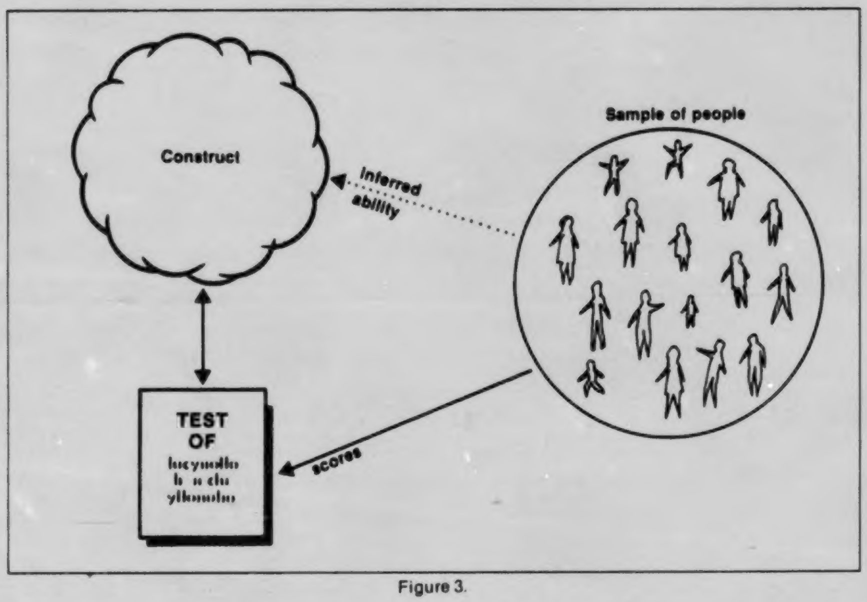


Figure 3.

Here we see a measurement procedure (a test) that has relations with several constructs—a strong one with Construct A, and weaker ones with Constructs B, C, and D. The problem of establishing the construct validity of a measurement procedure becomes one of determining what set of constructs are measured, and how important each of these constructs is in influencing the measurement outcome.

In the case of ability measurement, we shall assume (Figure 3) that both the construct and the measuring procedure pertain to inferred differences among individuals in some kind of ability or abilities.

The individuals are assumed to differ in some kind of ability, and the measurement procedure yields a score for each individual. These scores exhibit differences, or what psychometricians call *variance*. But even the inferring of individual differences in ability entails a further construct, namely, the construct of "ability" itself. There are many kinds of abilities (Figure 4), but all are exemplars of the general construct of *ability*.

Leaving aside the question of what categories of ability exist, and how many such categories there are, what do we mean by "ability" as a general logical construct? It is difficult to define the general concept of ability without

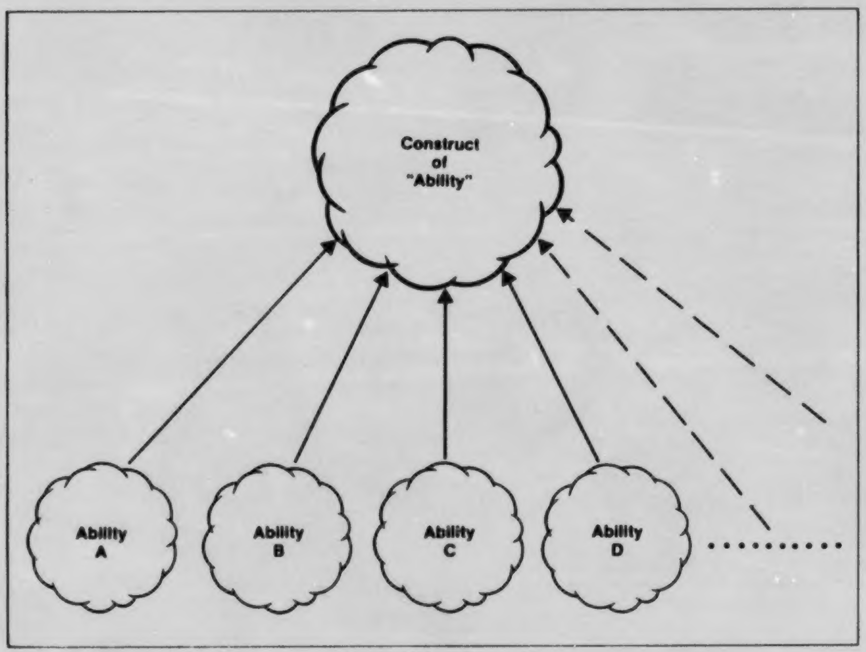


Figure 4.

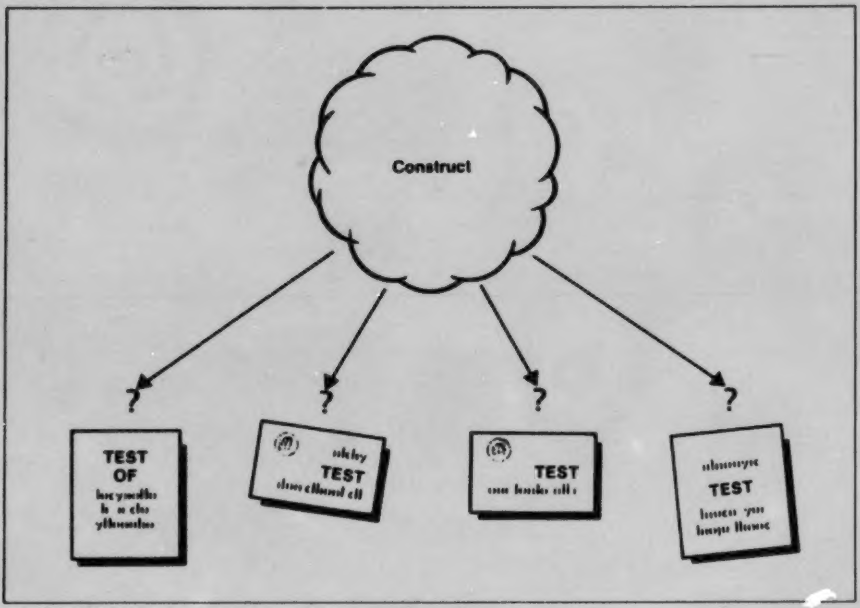


Figure 5.

being somehow tautological, because we are tempted to speak about ability in terms of what people "can" do. Let's try to be more explicit, avoiding tautology as much as possible. We could try something like this:

> An ability is any of one or more inferred properties or characteristics of an individual which makes it possible for the individual to perform more or less consistently at some level of success on some defined category of tasks or problems, or in some recognized class of situations.

You can easily find flaws in this definition, I am sure, but I will not try to improve on it here. One problem with this definition is that it does not begin to take account of the notion that increasing degrees of ability imply increasingly higher probability or amount of success for a given level of task difficulty, and that, as task difficulty increases, probability or amount of success decreases for a given level of ability. But amplifying the definition to take these notions into account would require consideration of what is meant by task difficulty, and that itself is a complicated matter.

Problems of construct validity seem to have arisen in two general types of situations. In one of these (Figure 5) the focus is on the construct.

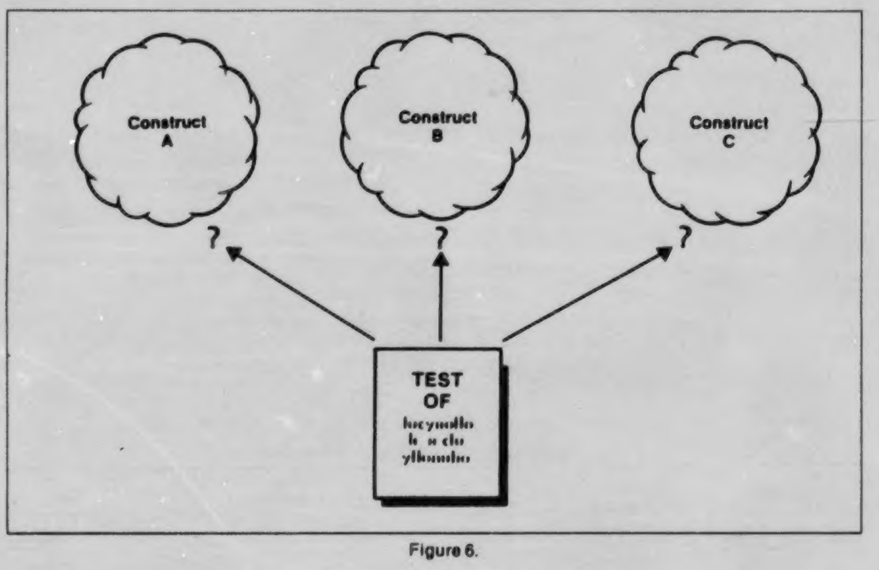


Figure 6.

One is given a construct, and the problem is to find some way of measuring it. I call this the *top-down* construct validity problem. The other situation (Figure 6) is the *bottom-up* problem: you are given a measurement procedure—one that you or somebody else has dreamed up, and maybe developed through sophisticated psychometric techniques—and the problem is to determine what construct or constructs, if any, this measurement procedure measures.

The measurement specialist is often confronted with the top-down problem. One is asked to provide measures of such constructs as "leadership ability," "success in job performance," "learning ability," or, perhaps, "intelligence," or "scholastic aptitude." Where do such constructs come from? Initially, they come from popular parlance and from the general experience of the culture. A construct does not necessarily have to be "an idea developed or 'constructed' as a work of informed, scientific imagination," as suggested by the 1974 APA-AERA-NCME *Standards for Educational and Psychological Tests* (1974, p. 29). A construct can be an idea that is present in the mind of the populace, developed over the experience of the human race, just as many concepts like "justice," "weight," or "comfort" have developed—none of them very scientific, at least at the start. Intelligence is one such concept, and scholastic aptitude and mental deficiency are still others, closely related. Further, I am quite sure that uninfluenced by any results of psychological research, popular wisdom would have it that people differ in such things as "mechanical ability," "spatial ability" (noticeable in people's difficulties in reading maps, for example), "musical ability," various kinds of athletic ability, and so forth. It would be interesting (and perhaps this has been done, though I'm not aware of it) to search through the literature of Western civilization for examples of constructs of various kinds of abilities. Plato's *Republic* recognized some of these abilities, and in *The Prince*, Machiavelli often considered how a ruler's effectiveness could be affected by his ability or wisdom. Constructs of generalized and specialized abilities have been with us for a long time, far back into ancient history.

Generalized constructs of intelligence and ability have been developed also in the context of sociological, economic, and historical analysis. For example, the sociologists Duncan, Featherman, and Duncan (1972) see intelligence as "a capability to function in the system of key roles in the society" (p. 78). The economist Raymond Boudon (1974) views intelligence as interacting with social status to produce differences in level of aspiration and thus generating inequality of educational opportunity. The historian Daniel Calhoun (1973) has written about "the intelligence of a people," that is, how intelligence was displayed in the American populace from the mid-18th century. As the culture changed, different kinds of abilities were drawn on—verbal intelligence in colonial preaching, spatial ability in the design of ships, and analytic ability in the building of bridges.

It was with such general concepts or constructs, derived from general public experience, that many test construction enterprises started. Binet's initial concern was with the construct of scholastic ability as seen in the different performances of normal and retarded children in school. George Bennett worked on the construct "mechanical ability," Carl Seashore set out to measure the construct "musical talent," and Norman Meier wanted to measure the construct "artistic ability."

But all that was long ago. In trying to operationalize various constructs, measurement specialists concocted an enormous array of tests and measurement procedures. Nowadays, many people may have forgotten what these tests were designed to do. The direction of thinking has reversed. Rather than asking how a certain construct can be measured, people now ask, what does such-and-such a test measure? What construct does it correspond to? Does it measure the construct that *I* (the person asking) have in mind? Such questions involve what I call the *bottom-up* problem in construct validity.

Some of the problems we have in communicating with the public about what tests measure arise, I think, because as measurement specialists we do not adequately recognize the role of generalized constructs of ability in the minds of the consumers, users, and "victims" of tests. When we try to be "scientific" about establishing the construct validity of a test (as the APA-AERA-NCME *Standards* urge us to be), we run the risk of making those clients confused and puzzled. Talking about tests in terms of "factors," "correlates," and "components" is all right among ourselves, but makes us all but unintelligible to the public at large. Frankly, this worries me. For all my own respect for scientific efforts to delineate the construct validity of our tests—efforts that I am about to describe and discuss—I think we must also respond in terms that test consumers can understand. The least we can do is to try to link "factors" and "components" to popularly recognized constructs of the type I have been alluding to. I believe this is possible, and that it can be done in a scientifically respectable way. I shall have more to say on this later.

Now let us look at the scientific approaches that have been made to assessing the construct validity of ability tests.

The Classical Approach

The APA-AERA-NCME *Standards* (1974) present a generally admirable, though brief, account of what may be called the "classical" view of establishing construct validity. Cronbach's chapter (1971) in Thorndike's *Educational Measurement, Second Edition* is an even more detailed account. According to the *Standards*, a construct is not only "an idea developed or 'constructed' as a work of informed, scientific imagination"; it must be "understood or inferred from its network of relationships" (p. 29) as established through correlational and experimental studies. They correctly point out that "it may be necessary to postulate several different constructs to account for the variance in any given set of test scores" and that "different constructs may be required to account for the variance in different tests of the same general type, or a given test may provide evidence relating to several constructs." The *Standards* might well have added that the construct validity of a test may vary over different types of populations, and over different types of testing conditions (as where the motivations of examinees might vary). Cronbach (1971) discusses correlational, experimental, and logical procedures for establishing construct validity. Perhaps because of the paucity of reported work using experimental and logical procedures, much of his discussion concerns correlational procedures, particularly factor analysis, with some attention to the use of nomological networks. He argues, I believe rightfully, against the extreme operationalism of those who would identify every single test or measurement procedure with its own construct, or of those who would define constructs only in terms of highly specific behaviors (like those of the "behavioral objectives" school of thought).

Having myself done a good deal of thinking and research within the factor-analytic tradition, I would like to amplify Cronbach's treatment of factor analysis as a methodology for construct validation, with particular attention to constructs of abilities. I want to discuss the overall motivation and strategy of the factor-analytic enterprise and attempt to show the continuity between it and newer approaches to the study of individual differences in ability, insofar as they relate to problems of construct validity.

As Cronbach remarks, factor analysis is a "treacherous technique." Unfortunately the history of factor analysis is replete with violent disputes over techniques, methods of interpretation, and the nature and meaning of the findings. Factor analysts themselves have contributed to creating the cloud of suspicion that still hangs over the technique in the minds of many psychologists. But factor analysis need not be a treacherous technique. Although there are many who still misuse factor analysis or are not fully aware of its pitfalls, experts in factor analysis have achieved substantial consensus concerning methodological and interpretative problems (Carroll, 1978). We are now in a much better position than previously to design adequate studies and produce replicable and reliable results. There is now a large body of well established knowledge that has come from such studies. Unfortunately, this

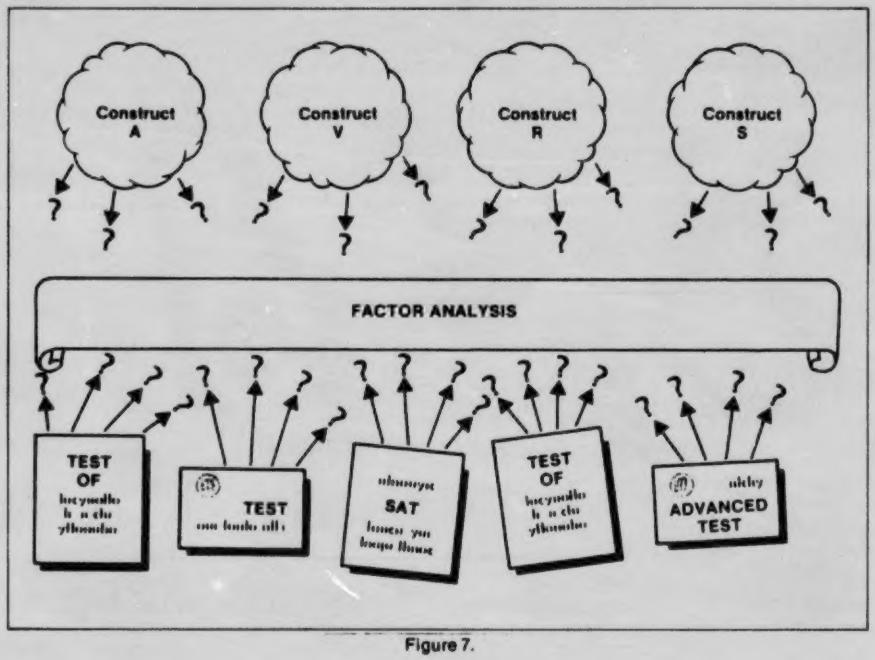
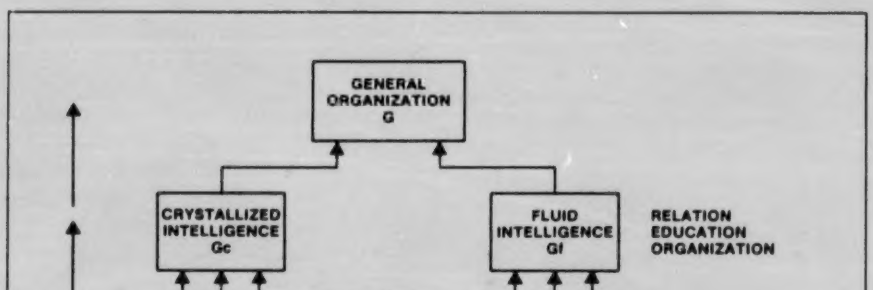


Figure 7.

knowledge is nowhere adequately summarized.

What is the factor-analytic enterprise all about? Essentially, it is an attempt to explore the variety of human individual differences in order to establish and precisely characterize a relatively small number of factors or constructs with which to describe those individual differences. As applied to the construct of ability, it surveys the ways in which human beings differ in their performance, as evidenced from various types of observations and measurements, to arrive at a more sharply delineated set of ability descriptions than those afforded by undisciplined general observation. By systematically varying the kinds of measurement procedures that are studied, and by conducting successively more disciplined and theoretically oriented studies, it offers the promise of identifying a system and structure of dimensions over which ability differences of all types may be interpreted.

Factor analysis operates somewhere in the middle of the region between constructs and actual measurement procedures (Figure 7), attempting to group measurement


GENERAL ORGANIZATION G
CRYSTALLIZED INTELLIGENCE Gc
FLUID INTELLIGENCE Gf
RELATION EDUCATION ORGANIZATION
ORGANIZATION VISUAL Gv
PERCEPTUAL SPEED Gs
ORGANIZATION AUDITORY Ga
PERCEPTUAL ORGANIZATION
Intelligence
SHORT-TERM ACQUISITION RETRIEVAL SAR
CORRECT DECISION SPEED CDS
QUIT DECISION SPEED QDS
TERTIARY STORAGE RETRIEVAL TSR
ASSOCIATIVE ORGANIZATION
SENSORY DETECTORS VISUAL Sdv
SENSORY DETECTORS AUDITORY Sda
SENSORY ORGANIZATION

N.B. Interpret Arrows Developmentally Rather Than As At-Time-Of-Measurement Determination

Figure 8. Schematic Representation of Intellectual Functions

procedures according to the constructs they measure, and at the same time refining the structuring of the "factors" that are inferred from these groupings. The inferred factors are often refinements of constructs that arise from common observation and experience. More importantly, they represent constructs that can be appealed to in establishing the construct validity of the measurements.

The grouping of factors appears to require the use of a hierarchical model whereby factors differ in their generality over different types of mental processes, knowledges, and tasks. Here (Figure 8) we see a hierarchical scheme proposed by John Horn (1979), one of the most active workers in the field of factor analysis.

At the top of the schema is a very general factor that, according to current theory, enters at least to some degree in nearly every type of cognitive task. Just below it are two somewhat less general factors, named "fluid" and "crystallized" intelligence, respectively, and at a still lower, that is, more restricted, level, stand factors of still less generality. (I will not try to explain what "fluid" and "crystallized" types of intelligence are.) Somewhere towards the bottom of the chart one would understand the presence of fairly narrow factors like those identified in Thurstone's researches—factors that Thurstone named Verbal, Space, Reasoning, Number, Perceptual Speed, and Word Fluency, for example.

In broad outlines, this type of structuring of the factors of cognitive abilities is well accepted. For a discussion of the so-called general factor of ability, I recommend Lloyd Humphreys' article in a recent issue of the Journal *Intelligence* (1979), on "The construct of general intelligence." Humphreys shows that a general factor of intelligence can be found to be involved in a very wide variety of cognitive tasks—even into tasks that are used to assess children's acquisition of the Piagetian concept of conservation.

Nevertheless, at a deeper theoretical level one can still suggest certain alternative interpretations of the general factor of intelligence. One of these is the notion that it is partly due to the operation of a speed or rate of work factor in many tests that are given under a time-limit or at least are given in situations where the examinee is implicitly placed under some time pressure. Another alternative explanation of the general factor, arising from the theoretical work of Atkinson and Raynor (1974), is that it may in part be due to the universal presence of motivational variables. Still another alternative interpretation of the general factor is that it relates to the subject's ability to handle the printed symbols that are so frequently involved in psychological tests—words, digits, etc. These alternative interpretations are open to experimental study, and one hopes that further inquiry will yield definitive answers as to how much each of the candidate hypotheses can be upheld.

To the extent that the identification and structuring of ability factors is converging toward established agreement, we can look upon it as yielding a sort of Mendeleevian periodic table of elements of cognitive abilities. We can quite possibly, now or at some time in the future, describe any cognitive test in terms of its factorial composition, that is, the extent to which it measures one or more of the factors or constructs identified in the table. For example, on the basis of various factor analytic studies we can specify the factorial composition of tests like the College Board SAT in terms of some combination of factors: a certain amount of the general factor "g," a certain amount of the so-called fluid intelligence factor, a considerable amount of the crystallized intelligence factor (which emphasizes the role of abilities acquired through education), a considerable amount of the verbal comprehension or verbal knowledge factor, a substantial component of general reasoning, and perhaps a small amount of the spatial perception or spatial reasoning factor that would show up particularly in the mathematical part of the test. The SAT thus has a rather complex factor composition, and its construct validity can be described only by appealing to a number of factor constructs. In effect, high scores on the SAT are attained by people who happen to be high in all of a number of abilities. We know also that the construct validity of standard intelligence tests like the Stanford-Binet IQ test and the Wechsler Adult Intelligence Scale has to be described in terms of a number of factorial constructs.

Various specialized aptitude tests, like mechanical aptitude tests, foreign language aptitude tests, clerical aptitude tests, and so forth, can be described in terms of certain combinations of more narrowly defined constructs. For example, clerical aptitude tests have a large component of what is known as the Perceptual Speed factor.

The factor analytic enterprise is far from being completed. It has been severely limited by the types of measurement procedures and observations that can readily be devised and administered to large numbers of individuals, and it has yet to study carefully the variety of cognitive tasks that are encountered in daily life, or in various nontest situations. In principle, nevertheless, factor-analytic methodology offers a way of explicating the range and classification of human abilities and setting them in correspondence with sharply defined constructs. While it cannot claim to provide a complete theory of cognitive ability, it does provide an initial set of data and findings that could be used in formulating such a theory.

I now turn to the consideration of certain newer techniques of establishing construct validity.

Newer Techniques of Establishing Construct Validity

As was pointed out in the review of recent research in human cognitive abilities that I published in the *Annual Review of Psychology*, Vol. 30 (Carroll & Maxwell, 1979), a number of information-processing psychologists like Earl Hunt, Robert Sternberg, Daniel Keating, and John Frederiksen have been putting their minds on the problem of endowing concepts of mental ability with greater meaning through the application of the techniques of experimental cognitive psychology. Their notion is to try to identify, in this way, the mental processes that people go through in performing ability and aptitude tests. If we can do this, we might have greater leverage in establishing the construct validity of such tests.

One of the first to take up this new line of investigation was Earl Hunt of the University of Washington. His strategy, which has been called the "cognitive correlates" approach by Pellegrino and Glaser (1979), has been to compare performances of high and low ability groups on various elementary cognitive tasks (Hunt, Lunneborg, & Lewis, 1975). The high and low ability groups are selected on the basis of SAT-type tests measuring verbal and quantitative reasoning abilities. The experimental tasks are selected from among the paradigms of current interest in cognitive psychology, such as choice reaction time, short-term memory tasks (e.g., the Saul Sternberg paradigm involving probed search of short-term memory), and the "Posner task" (in which the subject is required to report whether two alphabetic characters, presented visually in a tachistoscope, are the same or different with respect to either physical identity, in some trials, or name identity, in other trials). On the Posner task, for example, Hunt (1978) has found that high verbal ability subjects can make faster letter comparisons than low verbal ability subjects can, the advantage being more pronounced when the comparison is made on the basis of "name identity" (capital "A" having the same name as lower-case "a"). Does this tell us anything about the construct validity of the verbal abilities test? Probably not as much as we would like, because we know that the verbal ability test itself is factorially complex; we cannot tell from Hunt's findings what aspects of the verbal abilities test are correlated with ("measure the same thing as") the parameters of the experimental task. It is tempting to assume, as Hunt apparently does, that those parameters are correlated with the "g" factor aspect of the aptitude test, but the data admit of several alternative explanations. One alternative explanation that I find rather appealing is that there is an essential link between the mental speed aspect of the experimental task and the speededness aspect of the verbal aptitude test. Essentially, Hunt's work is an extension of correlational, factor-analytic methodology. With more careful design to exploit the advantages of that methodology, this type of work can possibly yield useful information about the construct validity of the verbal and quantitative aptitude tests that have been used as correlates of the experimental task data.

The research of Keating and Bobbitt (1978) and also that of Spiegel and Bryant (1978) extends Hunt's work along developmental lines. Like Hunt's, their work is an extension of correlational methodology, but they trace parameters of experimental tasks over a range of ages from childhood to adulthood. They show that these parameters change with age, and that they are correlated with mental growth as indicated by scores on such tests as Raven's Progressive Matrices. These findings, while of interest, do not at present exclude the possibility that the correlation depends on changes in the characteristic speed or rate of mental work parameters as children grow older, and not necessarily on changes in other kinds of cognitive growth. For there are various findings (e.g. Thissen, 1979) that suggest that the Raven test measures both mental speed and some other kinds of abilities.

One other investigator who has used correlational methods to claim that general intelligence is intrinsically bound up with parameters of certain elementary experimental tasks, like choice reaction time, is Arthur Jensen (1979). I have reanalyzed his results, however, and do not find clear support for his claims (Carroll, 1979). What I find in his data is at least two separate factors or constructs—one having to do with the speed or rate of information processing, and the other (uncorrelated with the first at least in Jensen's sample of university students) having to do with accuracy of response on tests of verbal knowledge and reasoning such as the Terman Concept Mastery test. In my reanalysis, the Raven Progressive Matrices score loads on both of those factors. The fact that the Raven score correlated to the extent of −.41 with the slope parameter of a choice reaction task led Jensen to conclude that mental speed was a property of the general intelligence construct, but this conclusion is far from compelling, because there is little justification for assuming the presence of a general factor in his data.

In historical perspective, there is little that is really new in the work that I have been describing. Back in 1944, in his factorial study of perception, Thurstone (1944) was already working with experimental tasks generally similar to those used by Hunt and others in more recent years. Thurstone even identified a reaction time factor, pretty much independent of his other factors, and some of his tests, for example a letter-comparison task, had certain features in common with the Posner task used by the contemporary investigators I have mentioned.

I have no intention of dismissing the recent work out of hand, but I want to point out that it often lacks the precision and methodological soundness that it ought to have, and that it has not yet provided clear support for its claims about the relation of elementary cognitive parameters to intelligence and abilities as measured by conventional tests.

There is another genre of recent work that may be more germane to the construct validity problem. This is research that looks directly at the aptitude tests themselves, on an item-by-item level. This is what Pellegrino and Glaser (1979) call the *cognitive components* approach. In 1976 (or actually in 1974, at a conference on the nature of intelligence), I pointed out (Carroll, 1976) that many items presented in conventional aptitude tests are highly similar, if not identical, to tasks studied in experimental cognitive psychology, and that the analysis of such tasks ought to help in understanding what is measured by aptitude and ability tests in which such tasks are to be found as items. In this connection I offered an introspective, somewhat speculative analysis of a sample of mental ability tests found in the ETS *Kit of Reference Tests for Cognitive Factors* (French, Ekstrom, & Price, 1963). My analysis suggested that those tasks contained information processing components that could be classified as processes, contents, and strategies. In a sense, I was pursuing a type of *logical* analysis of construct validity. What was lacking, of course, was the necessary empirical evidence for the information processing components.

Investigators have now started to remedy this lack. One of these is Robert Sternberg (1977), who is making a contribution to this symposium. Another investigator whose work falls in this category is John Frederiksen (1978), who has been able to delineate separate processes in reading behavior and to assess their importance in performance on conventional standardized tests of reading comprehension. According to his results, critical processes in reading behavior include "encoding multiletter units" and attaining "automaticity of articulation." Processes like these seem to explain a large part of the variance of the standardized reading tests. Nevertheless, these results are at least partly of a correlational nature, and their exact significance is difficult to assess. The results tell us that people who can quickly encode "multiletter units" tend to have higher reading comprehension scores than those who encode such units more slowly. But this cannot mean that poor reading comprehension is necessarily due to slow encoding processes in performing the test, for it is a standard rule that causation cannot be inferred from correlation. It may be that good reading comprehension and good multi-letter encoding abilities simply happen to co-occur in people. And even though Frederiksen was able to predict all the variance in one of his reading comprehension tests from the components he had isolated, this does not exclude the possibility that other components of the test could be found. For example, general vocabulary knowledge was not tested or otherwise studied in Frederiksen's research; yet, most reading specialists would argue that vocabulary knowledge is often implicated in reading comprehension tasks. Thus it would appear that even some of the newer ways of examining the components of mental tasks may fall short of providing the key information we need about the construct validity of tests. Nevertheless, I remain optimistic about the promise of the general approach.

Actually, it seems to me that there is one strategy, suggested by this work, that has been before our very eyes for years, without our having much used it in any deliberate or systematic way.

Construct Identification By Item Difficulty Analysis (CIBIDA)

What I propose is that we make better and more systematic use of item analysis and test score distribution data in assessing construct validity of an ability test. The method, for which I propose the acronym CIBIDA (Construct Identification by Item Difficulty Analysis), would best be applied to tests that are minimally speeded (so that nearly all examinees answer all items) and that have items that are reasonably homogeneous in character but vary quite radically in difficulty, all with reasonably high item-test biserial correlations. Essentially, the method is to compare items that differ in difficulty, trying to infer, in whatever way we can, what kinds of increased demands on mental processing, knowledge, and problem-solving strategy are made as the items become more difficult. I realize that Cronbach, at one point (1971, p. 483) in his chapter on test validation, cautions against accepting the statement of a test constructor who says, "If you want to know what my test means, look at the items or the operations." Cronbach states that "this thrusts the burden of construct validation onto the consumer, who will inevitably make inferences beyond the universe of situations representatively sampled by the test." Nevertheless, in the case of mental ability tests, it is often relatively easy to diagnose the mental operations, knowledge, or whatever, required by items and to make accurate inferences about the universe of items that are sampled. Thus I do not believe that the method I am proposing seriously incurs the dangers that Cronbach warns against.

Let me demonstrate the method with an easy case, namely a set of vocabulary items taken from an "open" form of the SAT. Through the kindness of Thomas Donlon of ETS I was able to secure not only the test itself (contained in a College Board publication called *Taking the SAT*, 1978), but also item analysis data.

					Estimated SFI values**	
	b*	p	Δ	r_{bis}	LEAD	choice
2. CONCEAL: (A) examine (B) recognize (C) expose (D) pronounce (E) arise	-2.70	.92	4.6	.50	41.9	41.9
4. STALE: (A) noticeable (B) fresh (C) dainty (D) moist (E) genuine	-2.50	.92	4.5	.58	42.7	59.9
6. STIFF: (A) limber (B) melted (C) succulent (D) twisted (E) silky	-2.60	.89	5.4	.53	53.7	38.4
5. DOUBTFUL: (A) practical (B) consistent (C) nonexistent (D) impervious (E) unquestionable	-1.65	.82	6.8	.53	47.0	40.4
3. EQUILIBRIUM: (A) opposition (B) insignificance (C) lack of freedom (D) lack of contact (E) lack of balance	-1.35	.83	6.7	.69	37.1	56.4
1. VENTURESOME: (A) lacking agility (B) lethal (C) fragile (D) timid (E) without significance	-0.82	.70	8.7	.66	36.1	45.3
7. PARTISAN: (A) commoner (B) neutral (C) unifier (D) ascetic (E) pacifist	+0.40	.44	11.8	.44	35.8	48.8
12. DISCREPANCY: (A) decision (B) attribute (C) restriction (D) clarification (E) concordance	+1.15	.25	14.2	.54	30.5	20.3
14. ELICIT: (A) leave intact (B) prove false (C) make acceptable (D) suppress (E) expel	+1.52	.20	15.0	.41	20	37.9
9. SOMBER: (A) pale (B) garish (C) fitful (D) extroverted (E) disordered	+1.60	.18	15.3	.46	43.0	30.6
13. WHET: (A) expire (B) heat (C) delay (D) slake (E) revive	+1.60	.20	15.0	.30	30.5	29.0
15. ENIGMATIC: (A) exceptional (B) explicable (C) exportable (D) expedient (E) exorbitant	+1.70	.19	15.1	.35	34.9	20
10. PAUCITY: (A) generosity (B) vacuity (C) excellence (D) variety (E) plenitude	+2.10	.16	15.7	.39	20	20.2
11. AMIABLE: (A) resigned (B) nervous (C) churlish (D) astute (E) obscure	+2.10	.13	16.3	.49	36.7	30.7
8. INFERNAL: (A) exquisite (B) frigid (C) ephemeral (D) mortal (E) celestial	+2.60	.17	15.5	.31	36.2	43.7

*b = value of z at which p = .5 (as estimated from item characteristic curve)

**SFI (Standard Frequency index) is on a logarithmic scale. Values as follows:

60 — 1 per 10,000 words
50 — 1 per 100,000 words
40 — 1 per 1,000,000 words
30 — 1 per 10,000,000 words
20 — 1 per 100,000,000 words

Figure 9. Vocabulary (Opposites) Items from Sample SAT (in order of difficulty)

In Figure 9 are shown 15 vocabulary items from Section 1 of this sample test. For these items, examinees are told to choose the "word or phrase (among the alternatives presented) that is most nearly *opposite* in meaning to the word in capital letters." Examinees are also told, "Since some of the questions require you to distinguish fine shades of meaning, consider all the choices before deciding which is best."

I have arranged the items, not in the order in which they are presented in the test, but in the order of their difficulty as I estimate those difficulties from the item characteristic curves that are plotted in Figure 10.

It is possibly of interest that these difficulties are not perfectly correlated with the *p* and delta values that are shown for the items. The item characteristic curves come from an item analysis sample of 1920 cases whose mean SAT score was 416 with a standard deviation of 110. The item analysis sample was divided into fifths, i.e., the lowest 20 percent, the next 20 percent, etc. Since the scores were normalized on the SAT scale, one can apply a formula for the mean score of a segment of the normal curve to arrive at the estimated mean SAT-V score for each fifth of the item analysis sample. These scores are shown along the baseline of the graph of the item characteristic curves; for each item, percent correct for each fifth of the item analysis sample is then plotted against these points. Note that for most of the more difficult items, the percents correct for low-scoring groups are well below chance levels; in fact, the curves tend to have a U-shaped concavity, probably because it is the low average group, at SAT-V = 358, that is most likely to be seduced into choosing an incorrect alternative. The really low scorers don't even have enough ability to be seduced in this way; they are the ones who are most likely to answer by guessing. This is a phenomenon that, I believe, is not taken account of in psychometric models of item response such as the Rasch model, or even the so-called three-parameter model.

Now, how can we look at these data to arrive at an estimate of the construct validity of these vocabulary items,

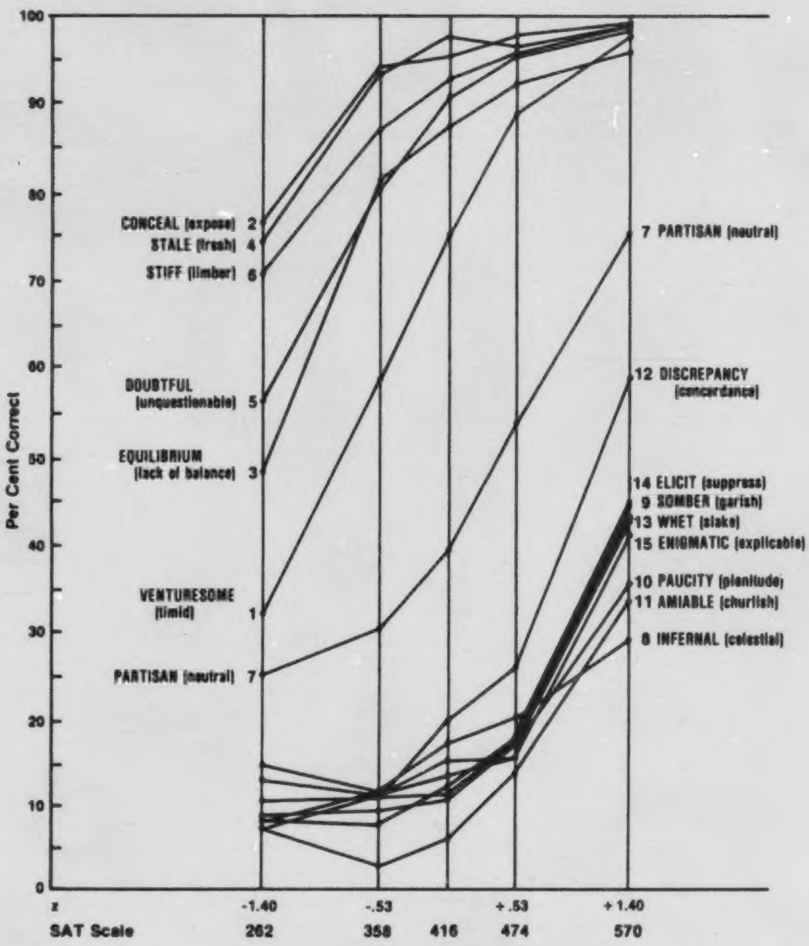


Figure 10. Item Characteristic Curves for Vocabulary (Opposites) Items from Section 1, Sample SAT Test

that is, at some statement about what they really measure? You may say that the answer is too simple, that obviously these items are a measure of extent of vocabulary knowledge. True, but let's consider this in more detail. First, what about mental processes? The examinees have to be able to read the items and recognize the words. Possibly some very low-scoring individuals have difficulty with that. Furthermore, the examinees have to know what an "opposite" is, and be alert to find opposites rather than synonyms. However, the item analysis data give very little evidence that any examinees had trouble with the concept "opposite." That is, the item analysis data give no evidence that any examinees were looking for synonyms rather than opposites, partly because very few if any of the distractors were actually synonyms. (One wonders why this was the case, in fact.) Therefore, we can conclude that these items are *not* measures of the ability of the examinees to pick opposites rather than synonyms. What gave them difficulty was the rarity and relative obscurity of the vocabulary words themselves, or of the particular meanings that were the basis of the keyed item choice.

I find it convenient to look at the items mainly in terms of the "lead" stimulus (the word in capital letters) and the keyed correct response (the word in lower-case letters). These words are shown directly on the graph of the item characteristic curves. The easiest item is one in which the examinees are supposed to recognize that *expose* is an opposite of CONCEAL. Hardly anybody in the high scoring group has any trouble with this, but about 24 per cent of the lowest scoring group got this item incorrect. The item analysis data show that their incorrect answers are more or less random, though more of them are seduced by "recognize" and "pronounce" than by the other alternatives.

The hardest item was the one in which *celestial* had to be chosen as an opposite to INFERNAL. The difficulty here was probably the relative obscurity of the meaning of INFERNAL as "pertaining to hell." One practically has to be a Dante scholar to recognize this meaning.

What makes this particular set of items interesting from the standpoint of construct validity is the fact that we are able to describe the items fairly accurately in terms of an objective measure of vocabulary difficulty. For this purpose we can consult various sources to get measures of word frequency in large counts of written English. In Fig. 9, I have listed the SFI (Standard Frequency Index) values found for the lead and the correct answer for each item, from the *American Heritage Word Frequency Book* (Carroll, Davies, & Richman, 1971). (I have had to make a few assumptions about words that could not be found in this source at all, or not in the exact grammatical form in which the words occur in these items). Computations show that item parameters can be predicted very well from these frequency values.

Overall, the item parameter that was best predicted was the ETS "delta" (Δ, the adjusted index of item difficulty). SFI's for the LEAD word and the *choice* word were intercorrelated .39 with each other, and to the extent of −.643 and −.687, respectively, with ETS Δ, yielding a multiple R of .799 ($F(2,12) = 10.67$, $p < .01$) both variables having significant beta-weights ($p < .05$). When the SFI variables were defined as the *lesser SFI* and the *greater SFI* for a given item, they intercorrelated .80, and had correlations of −.783 and −.738, respectively, with ETS Δ. This yielded a multiple R of .805 ($F(2,12) = 11.03$, $p < .01$), but with neither variable having a significant beta-weight. There was a suggestion that the test constructors consciously or unconsciously arranged the items according to the SFI of the rarer word, since this variable alone had a correlation of −.85 with item number.

Quite apart from this, we can use the item analysis data to say something useful about the meaning of different levels of SAT-V scores. Let's look at examinees with SAT-V scores of 570; this is near the mean SAT-V score of undergraduates at my own institution, incidentally—the University of North Carolina at Chapel Hill. These people can be expected to have no trouble with words like CONCEAL, STALE, STIFF, DOUBTFUL, EQUILIBRIUM, and VENTURESOME, and the keyed correct answers *expose*, *fresh*, *limber*, *unquestionable*, *lack of balance*, and *timid*. But I find it rather disturbing that they tend to have trouble with words like PARTISAN, DISCREPANCY, ELICIT, SOMBER, WHET, ENIGMATIC, PAUCITY, AMIABLE, and INFERNAL. Most of these words, and their paired correct answers, are fairly likely to occur in the texts students have to read at the college level. One's opinion of the vocabulary knowledge of people with SAT-V scores of 570 is somewhat deflated by these data.

If we look at the data for people with SAT-V scores of 262, the picture is pretty dismal. These are people who are totally lost with the more difficult items (that is, with the words contained in them), but they have trouble even with the easier items. The meanings of words like *limber*, *equilibrium*, *venturesome*, and *partisan* are unlikely to be familiar to them. These data tell a lot about the deficiencies of people with SAT-V scores as low as 262, and help us characterize the meaning of such a score.

Now let us turn to a slightly more difficult case in which we might use item analysis data to interpret the construct validity of a test.

In Figure 11 are shown a series of verbal analogy items

	b	p	Δ	r_{bis}
27. AMPLIFIER:HEAR:: (A) turntable:listen (B) typewriter:spell (C) platter:eat (D) camera:feel (E) microscope:see	-1.33	.82	6.8	.64
26. RIVER:CREEK:: (A) street:alley (B) hill:tunnel (C) path:sidewalk (D) hedge:fence (E) island:peninsula	-1.08	.76	7.9	.64
29. THIMBLE:FINGER:: (A) muzzle:snout (B) collar:neck (C) bracelet:wrist (D)helmet:head (E)yoke:shoulder	-1.08	.74	8.1	.60
28. HOMESTRETCH:RACE:: (A) finale:opera (B) goal:contest (C) boundary:journey (D) terminal:station (E) platform:campaign	-0.53	.63	9.6	.70
32. FLURRY:BLIZZARD:: (A) trickle:deluge (B) rapids:rock (C) lightning: cloudburst (D) spray:foam (E) mountain:summit	-0.15	.53	10.7	.59
33. SPATULA:ICING:: (A) trowel:mortar (B) knife:bread (C) sieve:flour (D) spoon:bowl (E) straw:milk	-0.10	.53	10.9	.68
30. STRUT:PROUD:: (A) stroll:eager (B) saunter:humble (C) scurry:pompous (D) shuffle:graceful (E) slink:furtive	+0.95	.25	14.2	.75
31. SWILL:SWINE:: (A) roe:fish (B) coop:poultry (C) mutton:sheep (D) pesticide:vermin (E) fodder:cattle	+1.40	.25	14.2	.43
35. AESTHETE:BEAUTY:: (A) enthusiast:cause (B) hunter:nature (C) administrator:government (D) advocate:legality (E) philanthropist: money	+1.70	.18	15.2	.52
34. BARREN:PRODUCTIVITY:: (A) torrid:warmth (B) innocuous:harm (C) aberrant:change (D) prodigal:reform (E) random:originality	+1.80	.16	15.8	.42

Figure 11. Verbal Analogies Items from Sample SAT
(in order of difficulty)

from this same sample SAT. The item characteristic curve data are plotted in Figure 12; the same item analysis sample is involved as before.

Note one thing about these items: if we formalize an analogy with the expression A : B : : C : D, the alternative choices represent a series of different possible C : D completions of the stem A : B. In terms of the mental processes described by Robert Sternberg (1977) as being involved in the solution of verbal analogies, examinees have not only to "encode" all the stimuli, and infer the relation between the A and B terms; they also have to find a choice that has a plausible "mapping" of a relation between the A and C terms, and one that correctly "applies" the rule inferred from the A and B terms. Let's go through this with the easiest item, number 27. From the A and B terms, AMPLIFIER and HEAR, the rule that must be inferred is that A is some *instrument* or *apparatus* that magnifies or increases the sensory sensitivity possible with a certain sensory modality indicated by the verb HEAR. Among the C terms, *turntable, typewriter, platter, camera,* and *microscope,* at least some are instruments or pieces of apparatus, but only *microscope* is one that permits increasing sensory sensitivity. Therefore it is "in the range of" A. Further, only the completion MICROSCOPE : SEE correctly applies the rule inferred from AMPLIFIER : HEAR.

Now, what is it about the more difficult items that gives examinees trouble? Let me make a prediction. As the items get more difficult, I would expect either or both of two things to happen. First, the vocabulary might get more difficult, in the same way as we saw with the straight vocabulary items. Second, the relations involved in the analogies might get more obscure and complex.

There is little evidence here that the examinees have difficulty with the concept and structure of an analogy per se. Even very low scoring individuals have a fairly good chance of passing a simple analogy like number 27, the one we have just looked at. The test might have included even easier analogies that would allow us to confirm this supposition, but probably such items were not included because they would not function as useful items in discriminating good and poor analogy solvers.

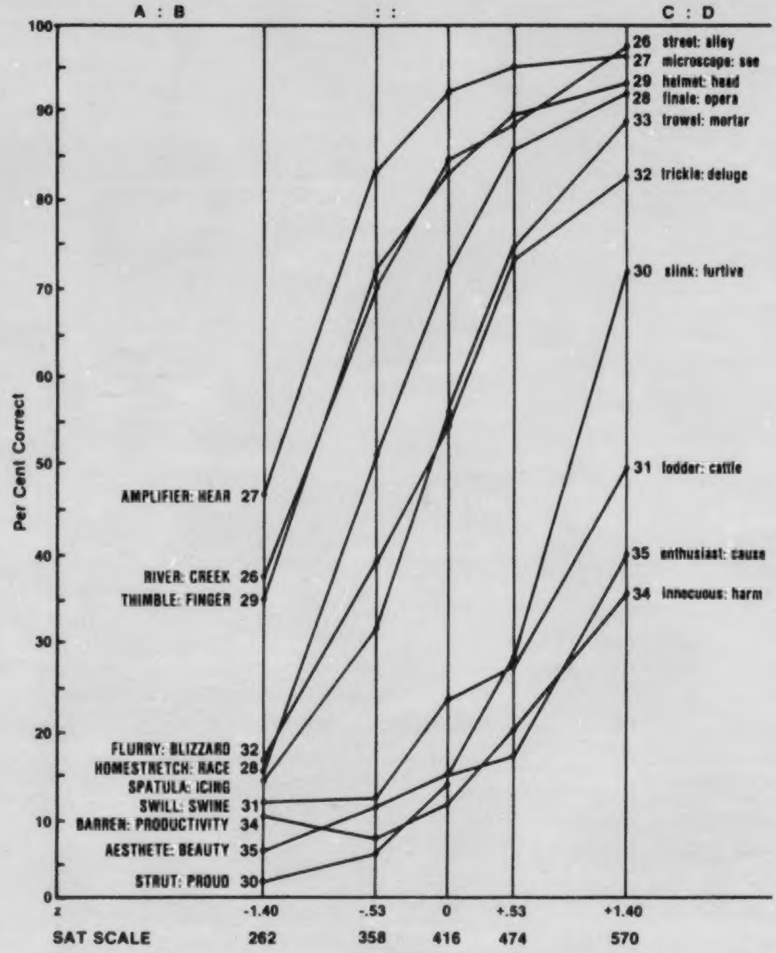


Figure 12. Item Characteristic Curves for Verbal Analogy Items from Sample SAT, Verbal Section 1

As evidence for the prediction that the difficult items can involve vocabulary problems, look at number 30, the hardest for low-scoring people, and somewhat difficult even for the highest-scoring group. There is nothing very difficult about the rule involved in the A : B relation, STRUT : PROUD. A is a verb for a characteristic gait or walk that conveys a feeling or demeanor described in the adjective, B. All the alternative choices also have verbs describing gaits, and adjectives describing feelings that might be conveyed by gaits. Further, all the C terms are "in the range of A." Only alternative (E), however, has a verb that correctly describes the demeanor conveyed by the adjective: *slink : furtive*. But *slink* and *furtive* are relatively rare or obscure words, as we can find from the *American Heritage Word Frequency Book*. The major difficulty here is undoubtedly of a vocabulary nature.

It is hard to find an item that does not have vocabulary problems, but possibly a good example is number 34. The words in the lead, BARREN : PRODUCTIVITY, are not particularly difficult words. The rule relating them is fairly complex: BARREN is an adjective, and PRODUCTIVITY is an abstract noun that conveys a property opposite to that of BARREN. The examinee has to find a choice that correctly exemplifies this relation. From the item analysis data (not shown here), we find that alternative C, *aberrant : change*, is rather tempting, as is also alternative (D), *prodigal : reform*, and without careful thought they might appear to exemplify the rule. Only alternative (B), *innocuous : harm* correctly exemplifies the rule, but it does so in a fairly subtle way. Admittedly, there may be a vocabulary difficulty; *innocuous* has an SFI value of 34.0. It is likely, however, that the complexity of the analogical relation is what gives most trouble, even for high scorers.

Finally let us consider items from the so-called "mathematical" section of the SAT. Part 2 of the Sample SAT contains 25 mathematical items that are somewhat heterogeneous in type. I have selected nine of them that exhibit distinctly different item characteristic curves; the actual items are given in Figure 13, and the item characteristic curves are shown in Figure 14. These data are from an item analysis sample of 1920 cases whose mean

1 If $2x + 2x + 2x = 12$, then $2x - 1 =$

(A) 2
(B) 3
(C) 4
(D) 5
(E) 6

2 If $\frac{1}{8}$ of a number is 3, what is $\frac{1}{3}$ of the number?

(A) 24 (B) 8 (C) 3 (D) 1

(E) It cannot be determined from the information given

5 $\left(\frac{2}{3} \div \frac{3}{4}\right) - \left(\frac{1}{9} \div \frac{1}{7}\right)$

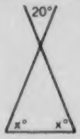
(A) $-\frac{9}{56}$

(B) 0

(C) $\frac{1}{9}$

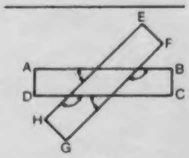
(D) $\frac{335}{630}$

(E) $\frac{8}{7}$



4 In the figure above, $x =$

(A) 20 (B) 40 (C) 70 (D) 80 (E) 160



13 In the figure above, if ABCD and EFGH are rectangles, what is the sum of the measures of the marked angles?

(A) 180°
(B) 270°
(C) 360°
(D) 540°
(E) It cannot be determined from the information given

17 If p, q, and r are integers and $\frac{q}{p}$ and $\frac{r}{q}$ are both integers greater than 1, which of the following is NOT an integer?

(A) $\frac{p}{r}$ (B) $\frac{r}{p}$ (C) $\frac{rp}{q}$ (D) $\frac{rq}{p}$ (E) $\frac{qr}{rp}$

20 If $\frac{7}{3}x = \frac{3}{7}y$ and $y \neq 0$, then $\frac{x}{y} =$

(A) $\frac{9}{49}$

(B) $\frac{3}{7}$

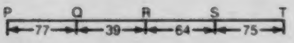
(C) 1

(D) $\frac{7}{3}$

(E) $\frac{49}{9}$

22 If the average of v and w is p and the average of x, y, and z is q, what is the average of v, w, x, y, and z in terms of p and q?

(A) $p + q$
(B) $\frac{p + q}{2}$
(C) $2p + 3q$
(D) $\frac{2p + 3q}{5}$
(E) $\frac{3p + 2q}{5}$



Note: Figure not drawn to scale.

24 In the figure above, if Q, R, and S are points on segment PT, the distance from the midpoint of QS to the midpoint of PT is

(A) 0 (B) 1 (C) 2 (D) 3 (E) 128

Figure 13. Items from Mathematical Section of Sample SAT Test (in order of difficulty)

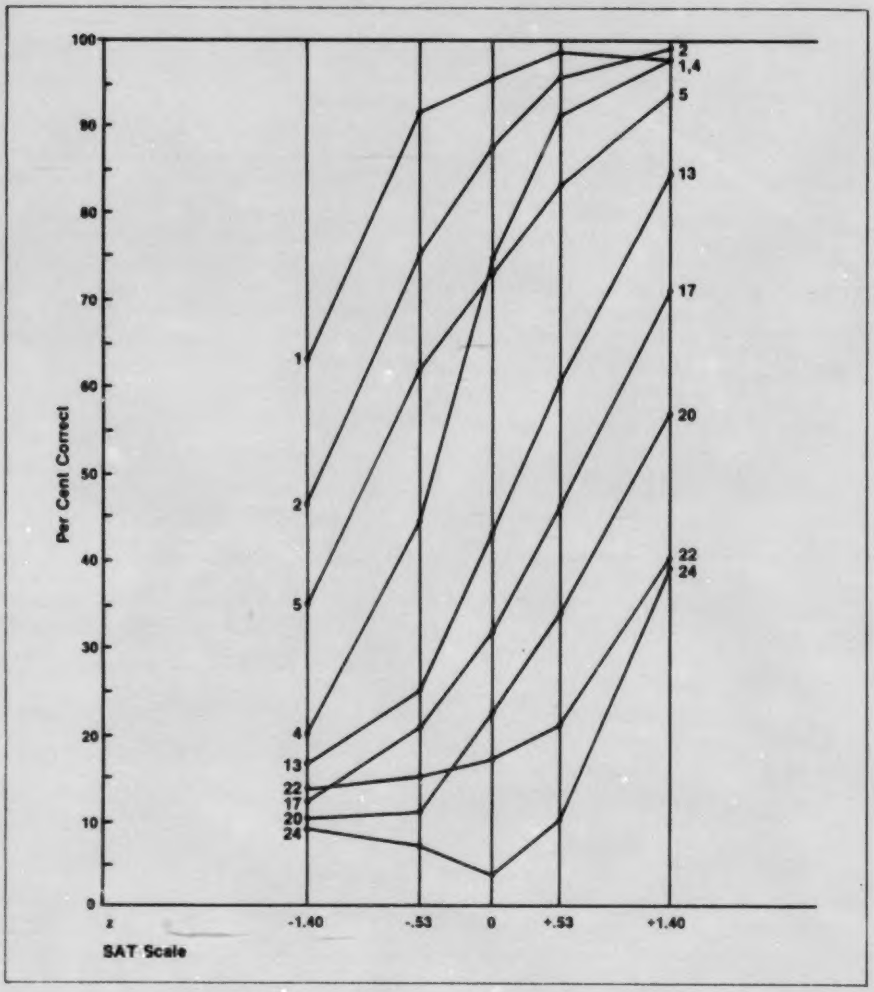


Figure 14. Item Characteristic Curves for Selected Items from Mathematical Section 2 of Sample SAT

SAT-M score was 455 with a standard deviation of 117. Using the same procedures as before, I show the estimated mean scores for successive 20 percent groups of the item analysis sample on the baseline of the graph for the item characteristic curves.

Again compare the characteristics of easy and hard items. Item 1 is very easy, as I think you will see by trying it right now. Nearly all high scorers get it correct, and only about 36 percent of the lowest scoring group get it wrong. It requires only very elementary algebraic manipulation. Items 2 and 5 are also easy; they require simple arithmetical operations that are "duck soup" for high scoring groups. Low scorers have a little trouble with them, but these items are still within their reach. Items 4 and 13 require knowledge of certain facts about angles; they are easy for high scorers, but quite difficult for low scorers, presumably because of their ignorance of the relevant geometrical facts or their inability to apply them even if known. Items 17, 20, and 22 are among the harder items, even for high scorers. They require subtle and involved algebraic manipulations, not easily accomplished without some paperwork. The hardest item is number 24, and why this is so difficult is somewhat of a puzzle. It has to do with a measurement operation that I can imagine a carpenter having to perform—finding the offset or difference between two midpoints. It could require some tedious arithmetic, but seems quite simple conceptually, and in fact can be solved by inspection. It was one of the last items in this subtest, yet was "reached" by 89 percent of the sample as a whole, and by 77 percent of the lowest scoring group. But 9.3 percent of the sample omitted this item—in fact, high scorers were more likely to dodge this item than low scorers; perhaps some high scorers perceived that it involved nothing but time-consuming arithmetic, and preferred to spend their time more profitably. Nevertheless, of those who tried the item, only the high scorers had any substantial success with it. Most people in the other groups were seduced into giving answer A (zero), apparently not noticing that the figure was not drawn to scale, and that the parts were drastically asymmetric.

From all this, I would say that construct validity of the SAT-M may be described as follows: It measures ability to perform fairly involved arithmetical and algebraic manipulations with care, accuracy, and speed; also, knowledge of simple geometrical facts about angles at intersections and in triangles gives the examinee an advantage. A person with a SAT-M score as high as 619 will have some difficulty with more involved problems; persons at the score level 291 will be totally lost with all but the easiest, most transparent items, and even these will be moderately difficult for them.

The constructs I have mentioned in this analysis of SAT items should not be difficult for the layman to understand. We have talked about vocabulary knowledge, ability to notice the relationships in verbal analogies, and the ability to perform arithmetic, algebraic, and geometric manipulations. We have *not* talked about "intelligence" or even "scholastic aptitude"; instead, we have been referring to the kinds of abilities that are directly involved in understanding the English language, understanding relationships among concepts, and understanding the manipulation of quantitative relationships. It is my belief (and practically certain knowledge) that if we could make similar analyses of many kinds of aptitude and ability tests, we could express their construct validity in terms that are as relatively simple and comprehensible as those I have employed here.

In short, there is no real mystery in the construct validity of ability tests, as long as we deal with it directly and concretely.

References

American Psychological Association. *Standards for educational and psychological tests*. Washington, D.C.: Author, 1974.

Atkinson, J. W., & Raynor, J. O. (Eds.) *Motivation and achievement*. Washington, D.C.: V. H. Winston, 1974.

Boudon, R. *Education, opportunity, and social inequality: Changing prospects in Western society*. New York: Wiley, 1974.

Calhoun, D. *The intelligence of a people*. Princeton: Princeton University Press, 1973.

Carroll, J. B. Psychometric tests as cognitive tasks: A new "Structure of intellect." In L. Resnick (Ed.), *The nature of intelligence*. Hillsdale, N.J.: Erlbaum, 1976. pp. 27-56.

Carroll, J. B. How shall we study individual differences in cognitive abilities?—Methodological and theoretical perspectives. *Intelligence*, 1978, *2*, 87-115.

Carroll, J. B. On the computation of the "g" factor. Paper presented at the meeting of the Society for Mathematical Psychology, Brown University, August, 1979.

Carroll, J. B., Davies, P., & Richman, B. *The American Heritage word frequency book*. Boston: Houghton Mifflin, 1971.

Carroll, J. B., & Maxwell, S. E. Individual differences in cognitive abilities. *Annual Review of Psychology*, 1979, *30*, 603-640.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement, Second edition*. Washington, D.C.: American Council on Education, 1971. pp. 443-507.

Duncan, O. D., Featherman, D. L., & Duncan, B. *Socioeconomic background and achievement*. New York: Seminar Press, 1972.

Frederiksen, J. R. *A chronometric study of component skills in reading*. Cambridge, Mass.: Bolt Beranek & Newman, 1978. (Report No. 3757.)

French, J. W., Ekstrom, R. B., & Price, L. A. *Manual and kit of reference tests for cognitive factors*. Princeton, N.J.: Educational Testing Service, 1963.

Horn, J. L. Statistical weaknesses of a unitary construct of intelligence. Paper presented at the convention of the American Psychological Association, New York City, September, 1979.

Humphreys, L. G. The construct of general intelligence. *Intelligence*, 1979, *3*, 105-120.

Hunt, E. Mechanics of verbal ability. *Psychological Review*, 1978, *85*, 109-130.

Hunt, E., Lunneborg, C., & Lewis, J. What does it mean to be high verbal? *Cognitive Psychology*, 1975, *7*, 194-227.

Jensen, A. R. *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology* (Tuskegee Institute), 1979, *2*(3), 16-29.

Keating, D. P., & Bobbitt, B. L. Individual and developmental differences in cognitive-processing components of mental ability. *Child Development*, 1978, *49*, 155-167.

Pellegrino, J. W., & Glaser, R. Cognitive correlates and components in the analysis of individual differences. *Intelligence*, 1979, *3*, 187-214.

Spiegel, M. R., & Bryant, N. D. Is speed of processing information related to intelligence and achievement? *Journal of Educational Psychology*, 1978, *70*, 904-910.

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, N.J.: Erlbaum, 1977.

Thissen, D. Latent trait scoring of timed ability tests. Paper presented at the Computerized Adaptive Testing Conference, Minneapolis, June 1979.

Thurstone, L. L. A factorial study of perception. *Psychometric Monograph*, No. 4, 1944.

DISCUSSION

Cronbach: I would like the record to show, since you referred to a point that I made in my chapter in Thorndike's *Educational Measurement*, that we are not really in disagreement. That is, the sophisticated individual who looks inside an item, with all the tools that you have used, is very different from the person who looks at the outside of an item and says, for example, "That looks as if it's measuring physics." I remember that when I was working on constructing a physics test, the item writers sent us items that they thought measured knowledge of physics, but in looking carefully at the item analysis data, it seemed that the items really measured difficulty in reading figures and graphs. I remember also the case of two Civil Service examination items that on their face measured ability to alphabetize, but since they had a correlation of zero with each other, they may well have been measuring something else, or nothing at all. So I think I would still caution against trying to interpret the construct validity of items by simply looking at them, although I agree that careful use of item analysis data, along lines you have suggested, may be valuable.

Jackson: I wonder if you would consider it good psychometric analysis to develop a notion of a "partial difficulty coefficient," in other words, to consider the difficulty level of an item against a number of facets. For example, one facet might be reading comprehensibility; another would have to do with relations between products; and so forth—all this being based, perhaps, on factor analysis. Would that increase the scope of this kind of analysis?

Carroll: Yes, I think it would. It would be nice to have item analysis data for different kinds of groups, groups that have been trained in different ways. For example, we could look at item analysis data for groups that have had geometry in high school versus groups that have not, or for groups that have had certain kinds of training in verbal skills as opposed to groups that have not had such training.

Tenopyr: Have you tried this?

Carroll: No, but I'd like to go back to some of my own work in verbal ability, foreign language aptitude, and other kinds of things, and try out these methods. And I'd recommend that other people try them out with their own data.

Tenopyr: I'm curious about spatial tests that apparently can be solved by different strategies.

Sternberg: I just wanted to say that I guess I agree basically with what you are doing, although I don't see it as a substitute for the cognitive components type of work. Rather, it seems to me, when you are trying to find basic sources of item difficulty, the question that still remains is, can one come up with a theory—in this case, a theory of how word meanings or relations are acquired, or how word meanings or relations are recognized in a given context? Given that this is done, a possible step would be to use the procedures you have suggested.

Carroll: I would agree that it would be nice to have a theory as an underpinning for my approach.

Novick: Looking at Figure 10, for the more difficult items, there is a drop in the average percent correct as one moves from a z-score of −1.40 to −.53. In your judgment, is that real or is it a question of relatively small samples?

Carroll: No, I think the effect is quite real. Each of the percentages is based on an N of about 384, which is large enough to give significant results. The effect is even clearer if you look at the trends for individual alternatives, particularly the distractors. People are really being seduced by these distractors. Yet among the really low scoring groups, there are people who seem not even to have the ability to be seduced, for their answers are more spread out among the alternatives, more or less on a chance basis.

Novick: Doesn't that fairly strongly suggest that the construct being measured by a particular item may be very different at different ability levels, and it becomes questionable to say that this test or this item measures this or that?

Carroll: Yes, you could interpret the data in that way. As I have suggested, construct validity might vary over different levels of ability, different kinds of samples or populations, and so forth. That is, there could be different sources of item difficulty for different groups. On the other hand, it is still meaningful, I think, to talk about a single construct of vocabulary knowledge being measured by these items, even if we recognize that there are different areas of vocabulary knowledge. Many years ago, E. B. Greene constructed his *Michigan Vocabulary Profile Test*, which gives scores in different areas of vocabulary—human relations, commerce, government, physical sciences, biological sciences, mathematics, fine arts, and sports. I once factor-analyzed data from the test and confirmed that one could measure different areas of vocabulary knowledge. But apparently the people who construct the SAT-Verbal test try to sample the different areas of vocabulary knowledge in order to come up with a summary score.

Messick: Rather than saying that different items reflect different constructs at different ability levels, it looks to me from your evidence that the same ability is being measured for high scoring subjects, over the full range of easy and hard items, whereas for low scoring subjects, it looks as if it's being measured just for the easy items. For hard items you need more evidence. They may be guessing, they may be sweating, falling apart,—doing all kinds of things. So the evidence about what you call "seduction" would be very informative.

Carroll: The general point I would make is that in interpreting the construct validity of a test, it would be very profitable to look at these kinds of item analysis data, including the pattern of responses to the distractors or wrong alternatives, much more carefully than we have in the past, and use that information in making inferences about what is being measured.

Hunter: Somehow that knack of looking at tests, and item analysis data, was lost to people in educational psychology about 25 or 30 years ago. I'm happy to see it being rediscovered.

Carroll: All of this goes back to the problem I had in giving a really good definition of what is meant by "ability." I think the notion of what happens to the probabilities of success at different levels of ability is crucial to the construct of ability. In order to measure a given ability, items have to reflect that model.

NORMAN FREDERIKSEN

RESEARCH MODELS FOR EXPLORING CONSTRUCTS

The typical test user probably knows very little about what is measured by a psychological test, and I'm not sure that the testing experts know very much more. We have been more interested in the technology of testing than in what goes on in the head of an examinee when he responds to a test item.

The lack of information about the cognitive processes involved in test taking was apparently obvious to Marion Richardson and John Stalnaker (1935) when they wrote "We wish . . . to point out that psychologists do not know what abilities are involved in procedures such as writing examinations."

In a memorandum addressed to "Risers to a Challenge," Bob Ebel wrote that, "We sometimes claim that different test items require different mental processes. . . I submit that we know practically nothing about these mediating processes, and certainly not enough to identify *different* processes or to distinguish one from another." That was in 1963, and I don't imagine he has changed his mind since then. A few weeks ago I noticed a full page ad for *Psychology Today* in the *New York Times*, in which Bob Sternberg is quoted as saying that psychologists know "almost nothing about what it is that [tests] have been measuring." But research over the last 15 or 20 years has brought new methodologies and new findings, so that we can now at least begin to describe what goes on in the head when one writes or chooses an answer to a test question.

Finding out what a test measures is the goal of construct validation. Construct validity was defined by Sam Messick (1975) as ". . . the process of marshaling evidence in the form of theoretically relevant empirical relations to support the inference that [a test score] has a particular meaning." He went on to say that ". . . test validation in the construct framework is integrated with hypotheses testing and with all of the philosophical and empirical means by which scientific theories are evaluated." In other words, a study of the construct validity of a test is equivalent to a scientific investigation of the cognitive processes involved in taking the test.

If we take seriously the notion that construct validation is equivalent to a scientific investigation, it follows that not only hypothesis testing is involved, but also the prior step of hypothesis generation. How does one go about formulating hypotheses regarding the psychological processes involved in taking a test?

In the next few minutes I will make some suggestions about how one might go about generating such hypotheses. Then I will offer some suggestions about how to test those hypotheses. Finally, I will describe briefly a study now in progress that incorporates a number of these suggestions and that illustrates several ways of interpreting construct validity data.

The Development of Hypotheses Regarding Processes Involved in Taking a Test

One source of hypotheses about what a test measures is to take the test yourself and introspect—an old-fashioned word and an old-fashioned method.

Let me give you a brief case history involving a test called Reversed Reading. It was chosen by Bill Ward and me to mark a factor called Cognitive Flexibility. We chose it on the basis of published evidence that it measured a factor called cognitive flexibility—the ability to change sets, as in Maier's (1930) pendulum problem or Duncker's (1945) candle problem.

The items of the test are sentences in which the letters in each word are printed in reverse order. The rationale is that there is a well established habit of reading from left to right, and since these sentences require reading words from right to left, it is necessary to break the well practiced set. Here is a sample item:

YREVE ERAUQS SAH RUOF SEDIS True False

Some of the findings of our study made me wonder if the interpretation of Reversed Reading in terms of changing sets was correct. I spent a little time taking the test, constructing other versions of the test, and trying them, and I now question the assumption that taking Reversed Reading items is merely a matter of remembering to start at the right in reading each word.

For a skilled reader, the cluster of symbols making up a familiar word is seen as a perceptual unit, not as a collection of letters that must be decoded. When the letters are reversed (ESICNOC, TRAP) it is almost impossible to see the words as a perceptual unit because the contour of the pattern is changed. However, if the individual letters are reversed as well as the order of the letters (ƧƎᗡIƧ, ƎƧIƆИOƆ), the contour is preserved and it is somewhat easier to see the word as a unit. So the Reversed Reading task is much more complicated than merely remembering to start each word at the right. The task requires one to revert to component processing methods such as a second grader might use in sounding out letters and varying pronunciation until something comes out sounding like a word that makes sense in the context of the sentence. Interpretation of Reversed Reading as a measure of cognitive flexibility appears to be an oversimplification.

All of this is intended merely to illustrate the use of introspective methods as a start toward the process of construct validation—developing testable hypotheses about what a test measures.

The "think-aloud" technique is another old method that can be used to develop hypotheses regarding what a test measures. The idea is merely to give a problem to an appropriate subject with instructions to try to solve it and to tell you everything he is thinking during the process. Whenever there is a period of silence, the subject is asked to report what he is thinking. Newell and Simon (1972) made extensive use of the method in their research on problem solving. Tape recordings and transcriptions were made from which they developed a theoretical analysis of the process typically used in solving problems of a particular kind. As everyone knows, they then wrote computer programs, such as the General Problem Solver, in which the theory was incorporated. The hypothesis about the problem-solving process was tested by comparing the performance of a computer with that of human subjects.

Other methods for generating hypotheses might involve such things as observing the examinee at work; studying the records of his performance, such as the scratch work done in the course of solving a mathematical problem or the notes made in preparation for writing an essay; and studying error patterns to find what wrong algorithms might have produced them.

Brown and Burton (in press) have made an extensive study of the error patterns of elementary school students in doing arithmetic problems, using a program called the BUGGY program because it can discover bugs in the algorithms. Teachers know some of these wrong algorithms. For example, a wrong algorithm in subtraction is to subtract the smaller number from the larger, regardless of which is on top; another is to borrow and then proceed as though borrowing had not occurred. The teacher who recognizes a particular wrong algorithm can take appropriate remedial action; but the teacher who fails to identify the wrong algorithm is likely to attribute the errors to carelessness. Brown finds, however, that students tend not to make careless errors. Instead, they are remarkably consistent with respect to the methods that produce the errors. He has identified many wrong algorithms that most teachers don't know about. Using a computer to discover the consistencies in behavior that produce answers, wrong or right, is thus another way of developing hypotheses about processes involved in taking a test. It could presumably be applied to areas other than mathematics if we had a sufficiently well-structured representation of the task environment.

In the past 15 or 20 years a considerable literature has been developed by cognitive scientists that deals with the processes involved in performing a wide variety of mental tasks, including such things as remembering, estimating, rotation of figures, problem solving, rule induction, and many others. Many of these mental operations are presumably elicited by tests, and a fruitful source of hypotheses would be the reports of the research in areas that seem relevant to the test in question. Some cognitive scientists—Earl Hunt and Jim Greeno, for example—have taken as their task the study of cognitive processes involved in taking tests of the sort that are commonly used in education.

One of the findings that such a survey of the literature would reveal is that there are individual differences in cognitive processes and that tests may measure different things for different people. John French (1965) did a study on this topic almost 15 years ago, a study that deserves to be better known than it is. He gave a battery of 15 tests—three tests to mark each of five cognitive factors—to 200 Princeton University students and Princeton High School students. He later gave other forms of the tests individually to the same subjects, with instructions to "think aloud." For most of the tests he was able to identify two different approaches or strategies used in taking the tests. For example, two subgroups were formed on the basis of whether an "analytic" or a "visualization" approach was used in solving Thurstone's Cubes test, a measure of spatial ability.

He then factored the intercorrelations of the 15 tests separately for each subgroup. For the subgroup using the visualization approach, the loading of the Cubes test on the spatial factor was .5, and for those using an analytic approach the loading was nearly zero (.07). Other loadings showed that for those who used the analytic method, the Cubes test measured inductive reasoning. Thus even the supposedly "factor-pure" tests may measure different abilities for different individuals.

I should also remind you of the importance of counter-hypotheses—hypotheses that you don't believe are true but that should be ruled out. For example, does an arithmetic reasoning test consisting of word problems measure reasoning or verbal comprehension? Does an anxiety inventory reflect anxiety or desirability response bias? Does a speed-of-reading test measure speed of reading or knowledge of the subject matter? And of course there is the strong possibility that there are individual differences in all these areas.

Testing Hypotheses About What a Test Measures

Those of us who were brought up in the psychometric tradition think first of using correlational methods to verify hypotheses about what a test measures. We have a long tradition of studying the validity of a test by computing its correlation with a "criterion," the criterion usually being something already in the file that has practical importance, such as freshman average grade. Such studies throw little light on the question of what the test is measuring.

Correlational information would be more useful for construct validation of a test if the other measures adequately represented hypotheses about what the test is measuring. At the present state of our knowledge about what tests measure, instruments that satisfy this requirement are hard to find; but if we choose measures wisely, we can by a series of bootstrap operations move closer to an understanding of what our test measures.

The correlational approach, then, should involve the selection of a set of measures to represent the various hypotheses about what a particular test measures, including measures representing counterhypotheses; obtaining data from an appropriate group of subjects; and computing the intercorrelations of all these variables.

One can then stare at the intercorrelation table, but there are of course better ways of analyzing intercorrelations. This is especially true since Karl Joreskog (1970) published his general method for analysis of covariance structures. A method of analysis that we have found valuable is to use maximum likelihood confirmatory factor analysis in obtaining extension loadings of the test in question on factors representing the hypotheses to be tested.

As an illustration of the method, Table 1 shows some results of an investigation aimed at studying the construct validity of a test called Formulating Hypotheses. This is one of several tests that Bill Ward and I (Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1978) developed for eliciting behavior related to scientific problem solving. Each Formulating Hypotheses item is a brief description of a research investigation in the area of behavioral science. Subjects were college seniors majoring in psychology. Their task was to read the description of the investigation, study a graph or table showing the results, read a statement of the major finding, and then write hypotheses, or possible explanations, of that finding—not only the hypotheses thought to be correct but also others that should be considered in interpreting the results or planning another study. In another form of the test, the subject was given a list of hypotheses from which he could *choose* hypotheses instead of thinking of the ideas for himself—a sort of multiple-choice form which is machine scorable.

The entries at the left below the double line in Table 1 describe the scores on the experimental tests. *FR* refers to the free response test and *MS* to the machine scorable form. *Best Quality* is a score representing the mean quality of the hypotheses the subject thought were his best. *Mean Quality* represents the mean quality of all responses, and *Highest Quality* the mean of those that were best according to our scoring system. *Number* is the number of hypotheses written (or chosen, in the case of the machine scorable test), *Unusual* is the number of ideas that occurred rarely, and *Unusual-High* is the number of hypotheses that were both unusual and of high quality—"creative" ideas, perhaps.

Table 1

Maximum Likelihood Factor Analysis of Cognitive Variables with Extension Loadings

	Vocab. I	Quantitative Reasoning II	Induction III	Logical Reasoning IV	Cognitive Flexibility V	Expressional Fluency VI	Ideational Fluency VII	Knowledge of Psychol. VIII
Vocabulary	.93	0	0	0	0	0	0	0
Word Meanings	.88	0	0	0	0	0	0	0
Math. Aptitude	0	.78	0	0	0	0	0	0
Necess. Arith. Opers.	0	.72	0	0	0	0	0	0
Resourceful Arith.	0	.85	0	0	0	0	0	0
Letter Sets	0	0	.71	0	0	0	0	0
Locations	0	0	.66	0	0	0	0	0
Figure Classif.	0	0	.68	0	0	0	0	0
Logical Diagrams	0	0	0	.65	0	0	0	0
Nonsense Syllogs.	0	0	0	.54	0	0	0	0
Inferences	0	0	0	.62	0	0	0	0
Reversed Reading	0	0	0	0	.84	0	0	0
Sign Changes	0	0	0	0	.64	0	0	0
Making Sentences	0	0	0	0	0	.54	0	0
Arranging Words	0	0	0	0	0	.69	0	0
Rewriting	0	0	0	0	0	.38	0	0
Pattern Interp.	0	0	0	0	0	0	.62	0
Diff. Uses	0	0	0	0	0	0	.77	0
Unexpected Results	0	0	0	0	0	0	.70	0
Adv. Psychol. 1	0	0	0	0	0	0	0	.80
Adv. Psychol. 2	0	0	0	0	0	0	0	.90
FR Best Quality	.23	.24	.30	.31	.26	.14	.04	.34
Mean Quality	.25	.25	.38	.41	.23	.17	.04	.45
Highest Quality	.20	.31	.44	.41	.26	.31	.18	.47
Number	.20	.31	.36	.39	.26	.44	.40	.36
Unusual	.10	.22	.26	.23	.20	.38	.36	.21
Unusual-High	.13	.23	.25	.18	.16	.24	.16	.28
MS Best Quality	.21	.20	.37	.30	.33	.18	.07	.47
Mean Quality	.12	.18	.18	.25	.22	.19	.04	.06
Highest Quality	.16	.29	.47	.46	.32	.29	.15	.47
Number	.08	.16	.29	.22	09	.07	.06	.39
Unusual	.04	.06	.16	.11	-.01	.01	.07	.34
Unusual High	.14	.25	.30	.29	.14	.08	.07	.46
Rho	.22	.34	.40	.49	.32	.30	.12	.35

Above the double line in Table 1 are listed the names of 21 tests that were selected to represent eight cognitive factors. The factors represent eight hypotheses regarding abilities that we thought might influence performance on one or the other of the two forms of Formulating Hypotheses. The two fluency factors were included because we expected that fluency would be required for the free-response form but not the multiple-choice form. Knowledge of psychology is represented by a factor formed by the two subtest scores from the GRE Advanced Psychology Test.

The maximum likelihood confirmatory factor analysis program finds the best fit of the data to a factor model provided by the experimenter. We chose a simple-structure model which allowed a test to load only on one factor, as specified by the nonzero entries above the double line in the table. The data were found to fit the model quite well.

The entries below the double line are the extension loadings of scores from the two forms of the experimental test; these loadings may be interpreted as correlations with the factors. (Since they are based on the factor structure rather than factor pattern, the correlations do not represent the unique contributions of the factors to variance on scores.) The correlations show that knowledge of psychology is an important resource in taking both forms of Formulating Hypotheses and for both quality scores and number scores. Similarly, reasoning, particularly induction, has an important role for both forms and both kinds of scores. The major difference is that the fluency factors are related to number scores only for the free-response form, as we anticipated. The results generally confirmed our hypotheses and thus strengthened our judgments regarding the nature of the two forms of Formulating Hypotheses.

Our expectations were not fulfilled, however, in the case of the Cognitive Flexibility factor. Note that one of the two tests used to mark this factor is Reversed Reading. We chose to include Cognitive Flexibility because we liked the idea that ability to change set should have something to do with formulating a variety of good hypotheses. We expected that cognitive flexibility would be more strongly related to performance on the free-response form of FH than the machine-scorable form, and to number scores more than to quality scores. As you can see, we were only partly right. The correlations may be telling us more about "cognitive flexibility" than about Formulating Hypotheses.

Cronbach (1957) in his APA Presidential Address commented on the separation of psychology into two camps—those who used correlational methods and those who used experimental methods; and he called for a federation of the two groups, the use of experimental and correlational methods in the same study. After twenty years, the federation is beginning to happen.

A paper by MacLeod, Hunt, and Mathews (1978) illustrates how the two approaches were combined in a study of cognitive processes. They used the Chase and Clark (1972) sentence-picture comprehension task. The subject sees a sentence on a scope, such as "PLUS IS OVER STAR," and then a picture, which is either a plus sign over a star ($\overset{+}{*}$) or a star over a plus sign ($\overset{*}{+}$). The task is to indicate as quickly as possible whether or not the sentence is true with respect to the picture. The dependent variable is reaction time.

In the MacLeod, Hunt, and Mathews study, two components of the processes were separately timed. The first component was the time required to read and understand the sentence, and the second was the verification time—the time required to judge whether the sentence matched the picture. They found that two groups of subjects could be identified on the basis of the difference between the two latency measures. Members of one group had short reaction times for comprehending the sentence and long reaction times for verification; subjects in the other group had long reaction time for comprehension and short reaction time for verification.

Those in the first group were presumed to have used a linguistic strategy. In the comprehension period they formed an internal representation of the sentence. In the second period, they did three things: they (1) formed an internal representation of the picture, (2) converted it into a sentence, and (3) compared the two sentences. The other group used a different strategy. In the first period they did two things, according to the theory: (1) they formed an internal representation of the sentence and (2) converted it into a picture. In the second period, they had only two things to do: (1) form an internal representation of the picture and (2) compare the two pictures. The first strategy is apparently verbal, ending up with a comparison of sentences; and the second is a spatial one, ending up with a comparison of two pictures. Scores on verbal and spatial tests were used to verify the process hypothesis. It was found that verification time correlated with scores on a verbal test for the first group and with scores on a spatial test for the second. People with different cognitive abilities thus used quite different processing strategies. The experiment not only illustrates how experimental and correlational methods can be combined; it also provides another example of individual differences in cognitive processes. A spatial strategy is again found to be an alternative to another strategy.

A Study of Construct Validity for Medical School Selection Tests

Now I would like to describe briefly a study in progress that puts together some of the ideas I have discussed and provides an example of one way to conduct a validity study. The purpose of the study is to develop and validate new tests that might be used in admitting students to medical school; it is being conducted jointly by the National Board of Medical Examiners (NBME) and ETS for a consortium of medical schools, with financial support from the Robert Wood Johnson Foundation.

A while ago I disparaged the common practice of making selection tests first and then looking for some convenient "criterion" variable, such as freshman grade point average, to use in validation. In the medical school study, we decided to begin by constructing a criterion measure. Our NBME collaborators were responsible for this part of the project, and they produced a set of diagnostic problem solving tests of the sort called "patient management problems." Such a test simulates on paper a situation that might be encountered by, say, a resident on duty in the emergency ward of a hospital. At the outset, the examinee is given a small amount of information about an incoming patient and he is asked to indicate what diagnostic hypotheses first come to mind. Then he is asked what he wants to do next (e.g., interview the patient, do a physical examination, order laboratory tests) and to specify the items of information he will try to obtain. The information he seeks is provided, after which the examinee writes a new list of diagnostic hypotheses and embarks on a new cycle of information getting and hypothesis formulation. This continues until the examinee is ready to make his final diagnosis.

On the assumption that the best way to predict performance on a criterion test is to make another test just like it, we at ETS constructed a comparable set of experimental selection tests. The selection tests, however, were based on problems that did not require knowledge and experience in medicine. The problems ranged from a horticultural problem (why are Mrs. Jones' hemlock trees dying?) to a public health problem (why did the absence rate suddenly increase in certain schools in a city?).

The tests, both medical and nonmedical, yield quality and number scores like those obtained from Formulating Hypotheses. In addition, it is possible to obtain scores based on sequential events in solving a problem, such as the number of hypotheses that are dropped for good reasons, the number dropped for bad reasons, the number of ideas directly suggested by items of information, and the number of ideas apparently based on inferences from information given.

A third set of variables, which we call process variables, was included for the purpose of investigating construct validity; they included tests to mark cognitive factors and personality variables. In addition, other information was obtained such as MCAT scores, scores on tests of medical knowledge, grades, and ratings.

Criterion tests, experimental selection tests, and the process variables were all administered to fourth-year medical students. The selection tests and the process variables were administered to entering first-year medical students. The criterion tests only were given to a small number of third-year residents. The data generated from all this make possible at least nine approaches to assessing validity:

1. Using data for fourth-year students, we can obtain correlations between the selection tests and the criterion tests. If we are willing to assume that the criterion tests are valid, these correlations will reflect concurrent validity of the selection tests.
2. Again assuming validity of the criterion tests, we will in the future be able to obtain evidence of predictive validity of the selection tests by a longitudinal study involving the first-year students.
3. Using the maximum likelihood confirmatory factor analysis with fourth-year student data, we can obtain information relevant to construct validity of the criterion measures by examining their relationships to the process variables—the measures of medical knowledge, cognitive abilities, personality, ratings, and so on. To the extent that the correlations are consistent with reasonable hypotheses about diagnostic problem solving, our confidence in the criterion measures may increase. In other words, their construct validity will to some degree be established.
4. In exactly the same way, using the same process variables except for the tests of medical knowledge, we can look for relationships supporting the construct validity of the selection tests. This can be done for both first- and fourth-year students.
5. Similarly, it will be possible to investigate the construct validity of conventional tests, such as the MCAT and the NBME achievement tests, and other items of information used in admissions, such as undergraduate grade-point average.
6. Since our theoretical expectations are the same for selection tests and criterion tests (except for knowledge of medicine), we can, using fourth-year student data, see how similar the selection and criterion tests are with respect to their relationships to the process variables. If the patterns of relationships are very similar, we can conclude that selection tests and criterion tests measure essentially the same construct. This is a kind of validity that hasn't been named, so far as I know. Call it "construct similarity."
7. Construct similarity can also be investigated in another sense: we can compare first-year and fourth-year students with respect to relationships of the selection test to process variables, to see if the test measures the same thing for students who differ in amount of medical training.

8. Construct similarity of the selection tests and the criterion tests with any other measure can be carried out by including the other measure as an extension variable in the confirmatory factor analysis. For example, the MCAT scores can be compared with scores from the experimental selection test with respect to their correlations with process variables.
9. Finally, score distributions obtained from fourth-year students and third-year residents can be compared to see if scores tend to change in ways that would be predicted on the basis of three years of experience in the practice of medicine.

In addition to the correlational studies, we plan to carry out a step that should have been first—to investigate by more direct methods the cognitive processes involved in taking the problem-solving tests.

First, I will have to explain in a bit more detail how we score tests such as Formulating Hypotheses and the problem-solving tests used in the medical school selection study. We call the method *category scoring*. The categories are the elements of a classification of the ideas written by examinees. This classification is used to simplify scoring of the free-response tests. Members of a panel of judges are asked to make judgments of the quality of each category and a scale value is assigned. The scale values can be stored in a computer, along with other information such as the frequency of occurrence of the ideas in each category. The scorer has only to assign each response written by a subject to one of the categories. When we give the computer the category numbers, it can generate scores of the kind I described earlier.

Our plan is to use these scoring categories in research aimed at finding out in more detail what is going on in the head when a subject works his way through the nonmedical problems used in our medical school selection research. The categories themselves can be classified and arranged to form a hierarchical structure, usually a multiple tree structure. Each tree represents a particular class of hypotheses; in the case of the dying hemlock trees, the classes of hypotheses are such things as soil, water, chemicals, and competition with other plants. The branches and twigs in the tree structure represent levels of specificity; for example, *chemicals* breaks down into fertilizer, lime, herbicides, and so on, and, at the twig level into things like "too much fertilizer," or "weed killer on the foliage." This graphic representation of the problem can be used as an encoding form for use in studies of how such ill-structured problems are solved.

We plan to administer the tests individually, using the "think aloud" technique to obtain information about what the subject is thinking. The form will be used to encode the behaviors of the subjects. The completed record, supplemented by transcripts of the oral reports of the subject, should provide a way to investigate how a given subject went about solving a problem; whether his hypotheses are scattered over all the trees or concentrated in one or two; whether his hypotheses tend to be "trunks" or "twigs" of the trees; and whether he tends to use items of information to make strong inferences that can eliminate many possibilities or merely to respond by association. We expect to find rather wide individual differences in the strategies employed, and we will be interested in looking for personal characteristics that are related to those differences.

Most of the research on problem solving has dealt with well structured problems, those where a small number of logical or arithmetical operations must be applied repeatedly in dealing with a relatively small amount of information. Creation of a production system that would simulate the behavior of a subject solving such a problem is largely one of determining which operations from a well-defined set of operations are employed and in what order. Our ill-structured problems entail a much larger amount of information, not all of which is present at the outset and not all of which is relevant, and the methods employed by subjects may be idiosyncratic and associational rather than logical. Therefore we expect to find a wider range of individual differences.

We hope also to carry out some experimental investigations to test certain aspects of the process models we expect to develop. For example, we want to compare groups of subjects who differ in the amount of knowledge relevant to a particular problem—for example, knowledge and experience in horticulture, in the case of the hemlocks problem. We may also want to compare subjects who have and those who have not had experience in a kind of problem solving in which certain strategies or heuristics are likely to be developed, such as in electronic trouble-shooting, solving diagnostic problems in medicine or playing chess. Such studies are intriguing because the encoding form can be thought of as a representation of the task environment, as the term is used by Newell and Simon (1972), and the encoded records of a problem-solver's responses can be used to infer how his problem space—his conception of the task environment—develops as he works his way through an ill-structured problem.

This brief description of studies under way and planned illustrates how a number of approaches to construct validation of a test can be employed in a study which is aimed at a very practical purpose, in this case, selection of medical school students. The results may make the product more useful for its intended purpose, and in addition contributions may be made to scientific understanding and to the development of cognitive theory.

References

Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in basic mathematical skills. *In Testing, teaching and learning: Report of a Conference on Research on Testing*. Washington, D.C.: National Institute of Education, in press.

Chase, W. G. & Clark, H. H. Mental operations in the comparison of sentences and pictures. In L. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley, 1972.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, *12*, 671-684.

Duncker, K. On problem solving. *Psychological Monographs*, 1945, *58* (Whole No. 270)

Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement*, 1978, *2*, 1-24.

French, J. W. The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement*, 1965, *25*, 9-28.

Joreskog, K. G. A general method for analysis of covariance structures. *Biometrika*, 1970, *57*, 23951.

MacLeod, C. M., Hunt, E. B., & Mathews, N. N. Individual differences in the verification of sentence-picture relationships. *Journal of Verbal Learning and Verbal Behavior*, 1978, *17*, 493-507.

Maier, N. R. F. Reasoning in humans. 1. On direction. *Journal of Comparative Psychology*, 1930, *12*, 115-143.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Newell, A., & Simon, H. A. *Human problem solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Richardson, M. W., & Stalnaker, J. M. Comments on achievement examinations. *Journal of Educational Research*, 1935, *28*, 425-432.

Ward, W. C., Frederiksen, N., & Carlson, S. *Construct validity of free-response and machine-scorable versions of a test of scientific thinking* (RB 78-15). Princeton, N.J.: Educational Testing Service, 1978.

DISCUSSION

Prien: Throughout this morning and part of this afternoon the discussions and comments have focused on construct validity and the operations and assumptions as they have been addressed and discussed throughout the past twenty-plus years. The issue that concerns me is more specific to the concern of this colloquium, and that is, whether we are talking about the same thing when we think of the Guidelines and the issues of our concern. The assumption of the equivalence of constructs in our classical thinking and constructs in the world-of-work setting is what concerns me. I think that's a faulty assumption that's prevailed here today.

Are we imposing a structure of concepts or constructs on the world-of-work—jobs, settings—that is not warranted?

Messick: There is a different type of construct.

Lawshe: Eric and I overlap. Are we talking about two sides of the same coin or are we talking about two different coins? I wrote down two questions here. One, what validity strategies are acceptable for determining what construct the test measures, and/or, that it measures the construct it purports to? And two, what validity strategies are acceptable in determining whether or not performance on a test and performance on a job tap the same construct?

Now virtually all discussion of this today has been around number one, which Eric described as the classical concept of construct validity. Now, those of us who are field practitioners confront the Guidelines that say now it's all right to use construct validity to demonstrate the *job-relatedness* of a test. And we haven't gotten very close to that.

Frederiksen: Isn't that what we are doing in the medical school study? Medical diagnostic ability is certainly something that's central to what a doctor has to do. That can exist at a conceptual level. But if you are going to get numbers, you have to do something to make that operational.

What we did was to construct simulations of situations in which a doctor can respond to a patient, as when a patient comes in and says, I have a pain here. We constructed other simulations that could be selection tests; they are similar except that they don't require any knowledge of medicine. We also used a battery of process measures that represent a set of hypotheses about the abilities required in solving the problems. The notion of construct similarity is to see to what extent the two patterns of relationships with process variables resemble each other.

Prien: My question pertains to choice of constructs. What constructs do you start with? Where do you enter the domain or what domain are you entering? What if you don't get a similarity? How will you interpret your results? What if the two are not similar in your analysis?

If the criterion doesn't load on any of the factors, does that mean the construct doesn't exist?

Frederiksen: No, it would merely mean the selection test didn't measure the same thing that the criterion measure measured.

Horn: It's quite possible you may end up not finding any construct similarity and still be able to predict. Then you have a problem of figuring out what it means.

Tenopyr: It seems you have started by defining a construct that represents a significant segment of our culture. I can't help seeing the similarity with what I said this morning. That is where you start.

Frederiksen: We got the advice of a lot of doctors in setting up the study, and they show two areas of great importance in the practice of medicine: one is diagnostic problem solving and the other is interpersonal relations.

Hunter: The way I would want to think about the question is this: what is being called for is an analysis of the criterion domain that's comparable in some respects to the analysis of the test domain, so that when you give us a table like this, instead of trying to give us some validity down at the bottom in accordance with these constructs that represent the top, you would give us a table in which the constructs would have been derived from known analysis of what medicos do, which might include a number of kinds of things which are not really abilities but have to do with, say, persistence in staying with the diagnosis, or maybe personality traits, and so on and so forth.

And in that sense, the constructs out there in the validity domain are really potential—not at all similar to the constructs that we are spending a lot of our time on in our theoretical domain.

Frederiksen: In the practical world, what kind of construct is of interest is the question of what is thought to be socially important by the people involved.

Hunter: Those are real constructs.

Prien: Would we not start with a concept of either effectiveness or worth in endeavors and take that as our point of departure?

Frederiksen: In medicine, I assume you would choose something that's judged to be most important. If you can't do everything at once, which is usually the case, you choose something that's very important from the standpoint of the welfare of patients who are going to be seen by doctors. It would be a lot more likely, I suppose, that patients tend to get well as a result of treatment rather than of the fact they have confidence in a physician.

Prien: It's an operational definition of purpose.

Tenopyr: Have you done anything relative to social skills at all?

Frederiksen: Yes, we do have a parallel study going on, an overlapping study, involving some of the same subjects, which involves the same idea—development of criterion tests and development of non-medical selection tests. The criterion tests center on the interview, the first interview between a doctor and a patient. We hired ten people to be simulators—we had five simulated patients and five simulated counselees. The criterion tests required doctors to interview people simulating patients; the non-medical tests involved interviews with people simulating counselees, people who came to the doctor to get advice on, say, "What medical school should I go to?" or "Should I play football?—a friend of mine had a head injury and I think it changed his personality"; and things like that. The same pattern in the research is involved, using personality variables as well as cognitive variables.

Tenopyr: Often social skills are not that socially important. If the doctor gets you well, he or she has fine social skills.

Frederiksen: There is some evidence that it is important. The patient who likes the doctor is more likely to follow the prescribed treatment, to take the medicine instead of throwing it down the sink, and so on. So it has some importance, other than keeping the income high for the physician.

Gulliksen: Is that known, or do they just guess?

Frederiksen: Studies have been done that show that effect.

Schmidt: Was there a reason why the medical simulation test wasn't used as the selection test, or some variation?

Frederiksen: Because the criterion test assumes medical knowledge.

Schmidt: What about the problems involving the hemlock trees and so on—are those on the test?

Frederiksen: On the non-medical test.

Farr: If I might return to test-taking behavior for a minute. You neglected to talk about Dick Shoales' work. His method is one of eye movement analysis while the subject takes test items. He's able to separate the examination of the question to understand the question, before one goes to the response alternatives, and uses various strategies to try to solve them. It's a very dissective method, and very informative.

I would also remark that Linn Cooper in doing some work on spatial rotation has found that strategies seem to play a very important part, over and above the so-called spatial and reasoning ability. It seems the data lead one to make the interpretation that a speed test in many senses is a very artificial test. Certain people who, for example, may lack spatial relations ability of some sort, visualization, rotational ability, cannot solve items at all in a speeded test situation. Given a little more time, using what Jack Carroll may have called minimum-speeded tests or no-speeded, one may bring in a different strategy, which involves a different ability, and get the problem right.

What this brings up with regard to construct validity is the question of using the general category of speeded tests. In the everyday world we aren't faced with the kind of time pressure the tests expose us to. I am talking about the average SAT. Student behavior in a course is not done on a split-second basis.

Guion: It seems to me there is something very important here. With the rather lofty language that we use when we are organizing our materials for an audience of psychologists, we keep talking about constructs, and when we get to the everyday world, we forget we are talking about constructs. You gave us a very nice example a few moments ago. I said this morning that the kind of reasoning we have to use to defend the test has the same logical form as what's been called construct validation, but it actually changes rather considerably when one gets down to everyday brass tacks.

You described a few moments ago that likeability, or the patient's reaction of confidence in the doctor, has been demonstrated by some kind of study, and it seems to be related to one small aspect of what is obviously important: Does the patient follow the treatment?

Now, if that were the assumption in the medical school when they wanted to use this test, then there is an obvious defense for using such a measure of likeability. This is an everyday construct; it has no high status in psychology. It almost certainly would not correlate with graduation from medical school. But it would be very easy to go into court and make a case, if this likeability measure is indeed a valid measure of likeability and, second, if there is a link in the network that says likeability is important to a physician's effectiveness, even though it's only one part of effectiveness.

These are the kinds of everyday things that we are dealing with when we try to deal with the demand for construct validation.

RICHARD H. McKILLIP and HILDA WING

APPLICATION OF A CONSTRUCT MODEL IN ASSESSMENT FOR EMPLOYMENT

The purpose of this paper is to describe the construct validity of Test 500, the written test portion of the Federal Government Professional and Administrative Career Examination (PACE). The continuing research and development effort spans several years. A fuller exposition of this work than can be presented here is contained in the publications listed at the end of this paper. We view Test 500 as construct valid because of the way it was developed, implemented, and documented. Although any two people in this colloquium, including the authors of this paper, might define construct validity in at least two different ways, the story of Test 500 provides an example, if not an operational definition, of construct validity.

The Origins of the PACE

A brief review of the background for the development of the PACE is in order because of the importance of considerations not applicable to other examinations of this type. Civil service competitive examinations must by law meet both merit system requirements and the requirements of the *Uniform Guidelines on Employee Selection Procedures* (USEEOC, USCSC, USDOJ, USDOL, USDOT, 1978). The Civil Service Reform Act of 1978 states:

> Recruitment should be from qualified individuals from appropriate sources in an endeavor to achieve a work force from all segments of society, and selection and advancement should be determined solely on the basis of relative ability, knowledge, and skills, after fair and open competition which assures that all receive equal opportunity. (5 USC 2301)

Implementation of this principle, of the similar merit principles it replaced, and of the *Uniform Guidelines* requires that certain conditions be met:

1. There must be a job analysis.
2. The knowledges, skills, abilities, and other worker characteristics (KSAO's) that underlie successful job performance must be identified.
3. Valid measures of the KSAO's must be developed.

Because of the large number of Federal occupations (over 2,000) and the large number of applicants (over 1,000,000 annually), the matter of matching applicants and the occupations for which they are qualified is extremely complex. In the centralized examining systems of recent years, occupations have been grouped for examining purposes on the basis of similarity of job requirements. In particular, since many professional, administrative, and technical jobs require the same kinds of knowledges, skills, abilities, and other worker characteristics, some 120 of these jobs have for many years been filled through a single examination.

The first such examination (the Junior Professional Assistant, 1939) selected applicants for 28 different occupations. It consisted of a review of applicant education and experience, a test of verbal comprehension, and a job knowledge test specific to the occupation applied for. Later, since maintaining the currency of the job knowledge tests was expensive, these subtests were replaced by requirements for specific experience or education. At the same time, subtests of quantitative and abstract reasoning abilities were added.

Several examining issues remained unresolved. First, the demands for such specialized preparation proved to be a barrier to liberal arts graduates and to others who, despite their lack of specific experience, had high levels of basic ability to learn and to perform the duties of the jobs. Second, many federal occupations have no specific experience requirements or have no industrial counterparts where people can gain experience. Internal revenue officer, customs inspector, and social insurance claims examiner are examples. Third, employment in the federal service often calls for adaptability to new techniques and attitudes, to advances in technology, and to shifts in social attitudes and practices.

The requirement of ability to learn new material appeared to relate to all three issues and also appeared to be a common requirement in many occupations. This consideration was, accordingly, the basis of the Federal Service Entrance Examination (FSEE), introduced in 1955. The FSEE had two basic requirements: appropriate experience and an adequate level of ability to deal with the complexity and difficulty of the journeyman or career-level demands of the occupations covered. The experience requirement could be met with three years of relevant work experience or with a bachelor's degree. A written test was used, at first assessing verbal, quantitative, and abstract reasoning abilities, but later dropping abstract reasoning when engineering occupations were dropped from coverage.

By 1970 it was clear that changes were again desirable. Technical advances made it feasible to introduce the differential weighting of test scores of large numbers of applicants. It also became feasible to measure additional abilities so as to reflect more accurately and comprehensively real, reasonable, and important qualifications underlying effective job behavior. It was also recognized that a systematic process was necessary to "determine the relative importance of the qualification requirements for the job in order to identify those of greatest relevance for measurement and, if appropriate, for later weighting when measures are combined" (USCSC 1972, p. 8). Existing methods of job analysis did not provide efficient identification of underlying abilities common to large numbers of occupations. For these reasons the United States Civil Service Commission decided in late 1972 to undertake the development of a new examination for the entry-level positions of professional, administrative, and technical federal occupations.

How the PACE Was Developed

Test development begins with the choice of a validation strategy. Of the three commonly recognized strategies, content validity appeared to be least relevant because of the absence of specific job knowledge requirements. Criterion-related validity seemed more applicable, but it could not be accepted as the sole strategy either, for two reasons. The first is the technical infeasibility of collecting enough job incumbents together to perform satisfactory statistical studies. For most of the occupations covered, there are a relatively few incumbents; these workers are often scattered among many geographical areas and many installations and may perform different mixes of the duties, depending upon the requirements of their employing agencies. Secondly, at the time this research was begun, this validation strategy had limited generalizability. Criterion-related studies in those occupations where data collection was feasible would say little if anything about the proper use of the written test for occupations where criterion-related studies were not feasible. A construct validity strategy, on the other hand, requires the hypothesizing of various psychological constructs important for performance both on the test and on the criterion; the development of measures of these constructs; and the gathering of evidence to support the relationship of the predictor and the criterion via the constructs and to support the explication of the constructs themselves. A series of criterion-related validity studies could provide much of this evidence.

To validate the new test, then, it was necessary to select the constructs, specify their assessment both in the job and in the written test, collect preliminary data supporting the identity of the constructs and the relationship of test to job via the constructs, and plan the subsequent research required for complete documentation.

The Objectives of PACE Development

More specifically, there were four research objectives to be met in order to support the claim of construct validity:

1. Identification of a set of cognitive abilities, or psychological constructs, required for successful performance in a broad range of professional, administrative, and technical occupations.

2. A systematic method of job analysis to provide an assessment of the relative importance of these abilities to duty performance.

3. Specifications for a written test which would assess these abilities so that differential weighting of abilities could be matched with differing job requirements.

4. Specification of the research required to provide additional support for the construct validity of the test.

To meet the first three of these objectives, a manageable number of occupations was chosen for intensive study. First, the duties of each occupation were listed. Then, the abilities underlying successful performance of these duties were derived. Next, the duties were rated for importance to the occupation and the abilities were rated for their importance to performance of each duty. These ratings were combined to determine the relative weight of each ability for each occupation. Finally, written tests were developed to measure the abilities.

Approximately 120 occupations came within the scope of coverage decided upon for this examination; all had been included in the Civil Service Commission's continuing program of occupational standards development and review. Because qualifications standards focus on minimum requirements, further job analyses were needed to identify the characteristics needed for superior performance. In order to reduce the job analysis problem to a manageable size, the researchers selected 27 of the most populous occupations, which accounted for over two-thirds of the hires during the early 1970's. The occupations include a wide range of activities, such as insurance claims examining, investigation, personnel management, writing, and administration. With few exceptions, it is unreasonable to require applicants for these occupations to show either specific academic preparation or work experience.

A large number of people participated in the research leading to the development of the test. Seventy-six separate groups of subject matter experts (SME's), a total of 1,241 persons, formed a sample of employees in senior-level and supervisory positions from the 27 occupations. They were chosen from 26 Federal agencies in locations throughout the country. In addition, a group of six experienced personnel research psychologists not otherwise associated with the study was selected to review the abilities in each of the 27 occupations.

The First Objective

Research objective No. 1, the identification of abilities, was accomplished in this way:

Established classification standards for the 27 occupations were reviewed in order to extract the major duties performed at the grade level occupied by the greatest number of employees, that is, the journeyman or full performance level. During this process the researchers educed a set of 31 knowledges, skills, abilities, and other worker characteristics (KSAO's) underlying successful performance in the occupations. These variables were then defined in terms of the behaviors associated with them, and, where appropriate, the definitions included examples of job situations in which the KSAO's were judged to be important. These variables, which are also described in the research literature, were to provide a framework for describing the occupations in terms of the attributes of successful performance, and were to provide a basis for selecting abilities to be measured by the written test. Each of the 1,241 subject matter experts (SME's) was asked to rate each of the 31 KSAO's for its overall importance to successful job performance.

Since the list of KSAO's and the ratings by the SME's were not sufficient to solve the problem of identifying testable abilities that would apply to all the occupations, the list and the ratings were compared with other evidence concerning abilities measurement. Attention was focused on cognitive abilities because of their known usefulness in predicting job performance, and because of regulatory and practical restrictions on measures of other KSAO's.

The 31 KSAO's were compared with the 59 factors listed in French's (1951) monograph. Item types described in that work form the basis of most of the abilities tests in use today. Six matches were found—a reasonable number considering that neither the set of KSAO's nor the French list had been designed to consist solely of cognitive abilities. The matches are listed in Table 1, together with the number of instances in which each is identified in the monograph. The KSAO *general reasoning ability* relates to the French factors of Deduction and Induction; the KSAO's of *arithmetic computation* and *quantitative reasoning* relate to French's Number factor. *Quantitative reasoning* is often found to load on Deduction and on Verbal Comprehension factors.

The six KSAO's for which there were reasonable matches with abilities identified by French were then analyzed for their relationship to job performance. They were: Verbal Ability, Judgment, Induction, Deduction, Number, and Memory. These abilities are well established in the literature, they appeared to be reasonably related to successful performance in the kinds of occupations under study, and they constituted a manageable number to be included in a written test that must be administered to large numbers of applicants.

Table 1

PACE Cognitive Abilities Matching
Factors Described by French

Cognitive Ability	Factor
General Reasoning	Deduction (37)
General Reasoning	Induction (9)
Judgment	Judgment (5)
Memory	Associative Memory (16)
Arithmetic Computation	Number (35)
Quantitative Reasoning	Number (35)
Reading Comprehension	Verbal Comprehension (46)

Note. Numbers in parentheses indicate the number of studies in which the factor was found.

The Second Objective

Meanwhile, Research Objective No. 2, the development of a job analysis method that would show the relative importance of the abilities, had been approached in this fashion:

Lists of duties had been derived from the classification standards. The SME's were asked to review the lists for their respective occupations. The first group to review an occupation was encouraged to amend the list of duties or to add to it as appropriate. Subsequent SME groups in the occupation did not change the list but were free to add duties to it. Then, each duty was rated—first, for its importance in the context of successful performance of the total job, and second, for the relative amount of time spent in its performance.

Next, the six personnel research psychologists participating in the study independently studied the classification standards and the duty lists as completed by the SME's. Working with the six abilities that had been selected for possible inclusion in the new written test, they rated each of these six abilities for its importance to each duty of each occupation, in strict accordance with the established definitions of these abilities and the occupational information.

The reliability of all three sets of ratings, as determined by standard procedures, was acceptably high, and nearly all reliability coefficients were statistically significant.

A pair of formulas was then developed to combine the two SME ratings of the duties for importance and for time spent with the psychologists' ratings of the duties for importance to each occupation. The first formula yielded values for the importance of each of the six abilities in each occupation. The second formula provided values for the relative importance of each ability within each occupation. Occupations were then grouped according to the similarity of their ability importance profiles.

In order to provide another view of the way the jobs tend to cluster according to the patterns of ability importance which they might share in common, factor analysis was carried out on the intercorrelation matrix of the ability importance values. This analysis was made across the six weighted constructs, that is, correlations in the matrix were based upon an N of 6. Factors were then interpreted using factor loadings, communalities, and percentages of variance accounted for by each factor. Five factors were extracted. All the variance was accounted for because there were five degrees of freedom and five factors. The factor analysis, therefore, could not properly be used to test hypotheses. The values were based on mean ratings and were, therefore, quite stable. Grouping occupations based on this factor analysis resulted in essentially the same groupings as that based on ability profiles.

It was considered important to determine whether the 31 KSAO's could logically be consolidated into fewer, more general groups, to corroborate the six constructs rated by the psychologists. A factor analysis, of the mean ratings of the 31 KSAO's, was done across the 27 jobs. Eight significant factors were extracted, accounting for 83.7 percent of the variance.

Eight of the first 11 KSAO's listed were noncognitive or personality variables while 17 of the last 20 were cognitive variables. The factor analysis confirms this distinction since the two sets of variables generally loaded on different factors. The factor analysis provided support for the abilities selected for inclusion in the test. The complex character, Memory, while not emerging as a factor in its own right, was present in four factors with a reasonably high loading on each.

The factor analysis pointed up the need for measurement of the ability to interact with people. The SME raters considered that many of the noncognitive characteristics reflect this one underlying factor. Parenthetically, a subsequent criterion-related validity study included as part of the criteria a duty entitled, "Interacting with and providing service to the public." The correlation of total PACE test score with this duty as measured by a work simulation exercise was .40.

Ability weights were now developed by applying the formulas and rounding the obtained values to integers. Seven distinct ability profiles, shown in Table 2, were required to describe the occupations in terms of relative ability importance. However, the kind of long term memory important for these occupations was confounded with other abilities in the factor analysis. Also, it was not possible to measure this kind of memory adequately with existing techniques for standardized testing. Therefore, two of the categories collapsed into one, so that there were finally six occupational categories: Category A, including 15 occupations, Category B, eight, and the remaining categories consisting of one occupation each. Five abilities, then, formed six ability profiles or occupational categories.

Lastly, it was necessary to assign each of the remaining PACE occupations to categories, since there are approximately 120 PACE occupations and the categories were developed on the basis of only 27 of them. Pending completion of a more comprehensive analysis, this assignment was accomplished by a senior occupational specialist thoroughly familar with PACE occupations, who evaluated each of the remaining occupations in the context of the five abilities.

Table 2

Rounded Weights of Abilities by Occupation

Series	Occupation	Deduction	Induction	Judgment	Memory	Number	Verbal Comprehension	Category
201	Personnel Management	2	2	2	2	1	2	A
967	Passport & Visa Examining	2	2	2	2	1	2	
993	Social Insurance Claims Examining	2	2	2	2	1	2	
1712	Training Instruction	2	2	2	2	1	2	
1810	General Investigating	2	2	2	2	1	2	
1811	Criminal Investigating	2	2	2	2	1	2	
1816	Immigration Inspection	2	2	2	2	1	2	
1890	Customs Inspection	2	2	2	2	1	2	
105	Social Insurance Administration	2	2	2	1	1	2	A
187	Social Services	2	2	2	1	1	2	
343	Management Analysis	2	2	2	1	1	2	
685	Public Health Program Specialist	2	2	2	1	1	2	
996	Veterans Claims Examining	2	2	2	1	1	2	
1081	Public Information Specialist	2	2	2	1	1	2	
1710	Education & Vocational Training	2	2	2	1	1	2	
110	Economist	2	2	2	1	2	2	B
180	Psychology	2	2	2	1	2	2	
526	Tax Technician	2	2	2	1	2	2	
560	Budget Administration	2	2	2	1	2	2	
570	Financial Institution Examining	2	2	2	1	2	2	
1102	Contract & Procurement	2	2	2	1	2	2	
1169	Internal Revenue Officer	2	2	2	1	2	2	
2001	General Supply	2	2	2	1	2	2	
334	Computer Specialist	2	2	2	1	2	1	C
962	Contact Representative	2	1	2	2	1	2	D
1082	Writing & Editing	2	1	2	2	1	3	E
1854	Alcohol, Tobacco, & Firearms Inspection	2	1	2	2	2	2	F

The Third Objective

Research Objective No. 3 was the specification of a test battery that would assess the abilities. The first component of this objective was the selection of item types; the second was the development of the test plan.

Identification of specific test question types to measure the abilities was deferred until the final determination of abilities to be tested was made. Selection of the question types was based on both technical and administrative considerations. Where feasible, two question types would be employed for each ability to reduce the effect of any method variance peculiar to specific question types. The French (1951) monograph and the *Manual for kit of reference tests for cognitive factors* (French, Ekstrom, & Price, 1963) served as the primary sources. Since the test was to be administered to some 200,000 persons per year, test question types had to be amenable to group administration with machine scoring of answer sheets. Accordingly, it was decided that questions would be of five-alternative multiple-choice format, but would otherwise follow the reference kit question types as closely as possible. Where deviations from the reference kit tests were to be made, care was to be taken that the mental processes required to answer the questions would conform to the mental processes required to answer questions in the corresponding reference kit tests.

The test specifications that were developed called for separate scores for each of the abilities (Verbal, Judgment, Induction, Deduction, Number), which would be weighted in different ways and then summed; these summed scores would be transmuted or transformed into employment ratings. In developing the test plan, it was judged that approximately 30 items would be required to provide adequate reliability in the assessment of each ability; and that two item types with 15 items each were needed. Existing resources provided two item types for each ability except Judgment, which would initially be measured only by comprehension items. A continuing program of item development would be necessary to provide the second item type for Judgment (Logical Order of Events) and to refine existing items.

An administrative requirement for alternate forms of the test meant that many items of each type were necessary. A formal equating system modeled on that of the Scholastic Aptitude Test was developed. The test battery was to include six parts, one for each of the five abilities plus an unscored part which could vary from one examinee to the next at a given test administration. This varying section could be an alternate form of one of the five others. All parts had the same time limit. Initially, each ability test was composed of 30 items and was allotted 35 minutes. The six parts required 210 minutes (3½ hours) of testing time.

The face validity of the test was carefully considered. To quote a colleague, "It's somewhat difficult to defend the selection of 'puissant' as a vocabulary item." Any materials used, such as paragraphs for reading comprehension or inference, or for tabular completion, would have to be in content areas that a federal employee in a PACE occupation might encounter on the job. Also, items were reviewed by women and by members of minority racial and ethnic groups, to eliminate items which might be offensive to any of these groups.

The Fourth Objective

The fourth and final objective was the specification of a research program that would provide additional support for the construct validity of the written test.

The setting up of such a program was difficult because the necessary operational steps are not well defined. Cronbach and Meehl's (1955) classic construct validity model, with nomological networks, has had little practical application. Procedures for testing hypotheses about constructs include logical analysis, correlational studies, and controlled experimental research.

What was done in the development of the PACE written test was similar to what is ordinarily done (or should be done) in developing employment tests. That is, generalizations were made on the basis of prior research, a test was developed and implemented based on these generalizations, and the necessary research involved the collection and analysis of data to evaluate the credibility of these generalizations. The claim for construct validity of the PACE lies in the research base of the generalizations about the cognitive processes required for performance in PACE jobs. The abilities selected have been shown to be related to successful performance in PACE-type jobs in hundreds of criterion-related validity studies (Reck, 1977); the item types selected to measure these abilities have been shown to be appropriate measures of them (French, 1951; French, Ekstrom, & Price, 1963).

The studies we have described so far report data pertaining to the appropriateness of the abilities to the jobs in question. A converging operation would be criterion-related validity studies in those selected PACE occupations where such studies were technically feasible. Schmidt and Hunter's (1977) model of validity generalization (Note 1) is an explicit and powerful methodology for combining data from many such studies, in order to assess the practical value of a test battery. The planning of our research program preceded the introduction of the Schmidt-Hunter model but in retrospect our strategy assumed the validity of validity generalization.

Studies directly supportive of the PACE test battery. Four criterion-related validity studies have been performed in support of the PACE test system. They are part of a planned series of studies in which the test scores of job incumbents are to be related to the scores of the same incumbents on specifically prepared measures of job performance. This work has already been described in a report of our colleague Brian O'Leary to the 1978 conference of the International Personnel Management Association Assessment Council (Note 2), so only the highlights are presented here.

The first four studies concerned three occupations whose duties are peculiar to the Federal government: Social insurance claims examiner (Trattner, Corts, van Rijn, & Outerbridge, 1977), internal revenue officer (O'Leary & Trattner, 1977), and customs inspector (Corts, Muldrow, & Outerbridge, 1977). For social insurance claims examiners, the PACE as a predictor of training success was studied as well as the PACE as a predictor of job performance (O'Leary, 1977).

In each study of job performance a task inventory was developed by asking journeymen to identify the tasks they performed and to indicate the relative amount of time spent on each.

From the task inventory, four measures of job performance were developed: A job information test (multiple-choice), for which items were developed by subject matter experts in the field; a work sample-case folder calling for decisions to be made, and, for customs inspectors, a special videotaped work simulation; a supervisory rating form tailored to the occupation and scaled to describe effective and ineffective performance; and a supervisory ranking form that contained the same description of job duties as the supervisory rating form but which had no scale points. (This measure was not used in the internal revenue officer study.)

The PACE and the criterion instruments were administered to 231 claims authorizers, 305 internal revenue officers, and 190 customs inspectors, at various locations throughout the U.S. Another group of 175 claims authorizers for whom training records were available took the test but not the criterion instruments. A total of 901 participants took these tests. Each participant was tested for approximately eight hours.

The correlations resulting from these studies are shown in Table 3. The total score on the PACE test was significantly related to job performance as measured by all the measures of job performance for the claims authorizer and internal revenue officer studies. For the customs inspector occupation, PACE scores were significantly related to performance on the job information test and the work sample but not the supervisory ratings and rankings. The pattern of validity coefficients was similar across occupations with a mean coefficient of .43. These results indicate that persons who score high on the PACE tend to perform better on the job.

Table 3

Correlation of Test 500 Scores with Criteria for Four Validity Studies

	Total Ability Weighted PACE Score	
Criteria and Occupations Studied	*r*	*P*
Job Information Test		
Claims Authorizers	.61	.001
Customs Inspectors	.65	.001
Internal Revenue Officers	.68	.001
Work Sample/Simulation		
Claims Authorizers	.36[a]	.001
Customs Inspectors	.56	.001
Internal Revenue Officers	.56	.001
Supervisory Ratings		
Claims Authorizers	.30	.001
Customs Inspectors	.06[b]	NS
Internal Revenue Officers	.25	.001
Supervisory Rankings		
Claims Authorizers	.31	.001
Customs Inspectors	.03[b]	NS
Training Performance Claims Authorizer Trainees	.51	.001
Mean Validity Coefficient	.43	(.49)[c]

a No correction for criterion unreliability because no estimate possible.

b No correction for criterion unreliability because bivariate correlation coefficient nonsignificant.

c Mean validity if customs inspector ratings and rankings not considered.

The correlation obtained between PACE and training success for claims examiners was .51, indicating that PACE is a valid predictor of training success.

Comparisons were also made of different procedures for weighting the subtests of the PACE. The construct weights which are now being used operationally produced validities that were essentially as high as those obtained by other weighting procedures.

Studies supporting the theoretical basis of the PACE. Of equivalent concern to the practical value of the test battery, evaluated by criterion-related validity studies, is its theoretical meaning. The factor-analytic literature currently provides the primary basis for this meaning, and our colleague, Lois Northrop, has reviewed this literature and related the PACE written test to this research (Northrop, 1976, 1977a, 1977b, 1977c, 1977d, 1977e). Many other studies have been done to document the characteristics of the written test battery. Magda Munoz-Colberg led the development of an item taxonomy, one that is multidimensional in that items are classifed by content and by logical structure, the logical structure being an elaboration of the definition of the ability to be measured (Munoz-Colberg, 1977, Munoz-Colberg & Nester, 1978; Munoz-Colberg & Nester, Note 3). Much practical work relating to the characteristics of the PACE test has been done in PRDC. It includes revision of the distractors in vocabulary items to reflect considerations of etymology as well as word class; tryout of different item types purporting to measure Judgment, with the conclusion that the two types originally selected performed as well as any other; and tryout of different item types, such as syllogisms, to evaluate the reasoning abilities of Induction and Deduction. Another project sought to improve the measurement of the Number ability. On early test forms the last one-third of the items had unacceptably high omit rates. Based on data from a number of unscored sections, the number of items was reduced from 30 to 20 and their character was simplified to reduce the effect of speededness.

Other PACE studies. Normative studies have been made of PACE competitors and selectees (Wing, 1976, 1978), as have studies describing the test performance of different racial, ethnic and sex groups (Wing, Note 4). Test fairness studies have not yet been technically feasible; item fairness studies are being considered. Reliability and equating studies have been performed on a continuing basis, the latter providing an opportunity for evaluating the effects of practice (Wing, Note 5). Intercorrelations among the test parts are reasonably consistent from one administration to the next (Wing, 1977).

An important concern is the assessment of handicapped examinees. From the beginning, forms of the written test battery were developed for the blind and the deaf, using appropriately selected items and item types so that these forms are parallel and equivalent to the regular competitive forms.

Other studies of existing data remain to be done. Studies of recompetition, in progress, will expand the knowledge about practice effects. The taxonomy needs a third dimension, that of the psychological processes required to answer the test questions. Wing has proposed an analysis of responses to the items of each type, to assist in determining the structure of the set (in the Garnerian sense) as a means of identifying the psychological processes required. It is possible that information-processing approaches, such as investigated by Carroll (1976), Hunt, Lunneborg, and Lewis (1975), Sternberg (1977), and others, will add to the PACE construct validity documentation.

Epilogue

The history of the PACE examination subsequent to the steps in its research and development is perhaps of interest. By the end of 1977, the four criterion-related studies had been completed, and the documentation of the constructs was written up in several reports. Work to improve the test continued, and research started on a new job analysis methodology for PACE occupations that would provide a greater generalizability and closer compliance with the Uniform Guidelines. Further criterion-related validity studies were to be conducted in conjunction with this effort. All evidence pointed to the conclusion that the written test was indeed predictive of job performance. But scientists are skeptical people, and nonscientist critics of testing in education and employment are more skeptical (if skeptical is the word) than the scientists. The Uniform Guidelines, for example, place the most stringent of validation requirements on employers who use selection procedures that have adverse impact. The inference of validity for the PACE test was strong; the further studies were expected to remove any vestiges of doubt.

Federal agencies were well satisfied with the caliber of employees selected from the top levels of the eligible registers. But, from the earliest use of the examination, there was concern that few blacks and Hispanics were being selected from the PACE. Applicant racial and ethnic data were not available; collection of such data had long been prohibited in order to protect minority group members from discrimination. In many locations, it was difficult to select women. Normative studies had shown that women, as a group, were adversely affected by veteran preference points added to the eligible list ratings of certain veterans (Wing, 1976).

The racial, ethnic, and sex makeup of the federal workforce in occupations covered by the PACE was known; approximately 20% minority and 42% female. These figures suggest that the "bottom line" of representation is reasonably met, despite the adverse effect of various components of the PACE. Federal agencies make extensive use of other avenues of entry into these occupations; such as promotion, upward mobility, and transfer. Nevertheless, there continued to be great concern about the examination, with particular emphasis on the effects of the written test.

In late 1975, a team from the General Accounting Office instituted a GAO audit on the job-relatedness and cultural bias of federal applications, examinations and standards (U.S. GAO, 1979). Racial data on applicants were obtained by matching test records to Social Security records. While the original objective of the audit was quite broad, the focus narrowed more and more on the PACE as the audit progressed.

From the first, the Director of the Personnel Research and Development Center explained the necessity for the audit team to acquire professional expertise in psychometrics. The team accepted this recommendation. The final report, which made many recommendations concerning additional research, did find substantial evidence of validity for the PACE written test, enough to meet the Uniform Guidelines requirements for continued use (U.S. GAO, 1979, p. 125). It also found adverse impact.

This finding poses a dilemma. The Civil Service Reform Act has dual objectives: a workforce from all segments of society, and selection on the basis of relative ability, knowledge, and skills. But the PACE test is consistent only with the latter objective. The Office of Personnel Management and other agencies continue to look for alternative valid selection procedures that do not have adverse impact.

If such procedures can be found, they will be embraced rapidly and exclusively. If they cannot, the objectives conflict, and either one must become supreme or some trade-off point determined. It does not appear that the objectives themselves will change in the near future. Certain kinds of tradeoffs, such as setting quotas and hiring the top scorers in each group, are politically inexpedient and of dubious legality, although they would have the greatest utility. Other kinds of tradeoffs may be politically and legally palatable, but may have reduced utility.

There is a growing awareness that valid selection procedures reflect differential group preparedness for education and employment, as well as differential individual preparedness. Consensus has not yet been achieved with regard to reconciling these group differences in preparedness with the objective of representativeness in the arenas for which the preparation is required. Clearer understanding of technical concepts such as construct validity can help to clarify the distinctions between the two objectives.

Notes

1. Schmidt, F. L., & Hunter, J. E. *Development of a general solution to the problem of validity generalization* (TS-78-1). Office of Personnel Management, Personnel Research and Development Center. August 1978.
2. O'Leary, B. S. An assessment of a construct validity paradigm. In R. H. McKillip (Chair), *Innovative test validation strategies*. Symposium presented at the Annual Conference of the International Personnel Management Association Assessment Council, Atlanta, Georgia. June 28, 1979.
3. Munoz-Colberg, M., & Nester, M. A. *Utilization of taxonomic and other construction criteria in the written test portion of the Professional and Administrative Career Examination* (PACE) (PRR-79-5). Office of Personnel Management, Personnel Research and Development Center, in preparation.
4. Wing, H. *Profiles of cognitive ability for different racial/ethnic and sex groups on a multiple abilities test battery* (TM-79-7). Office of Personnel Management, Personnel Research and Development Center, in press.
5. Wing, H. *Practice effects with traditional mental test items* (TM-79-9). Office of Personnel Management, Personnel Research and Development Center, in press.

References

Carroll, J. B. Psychometric tests as cognitive tasks: A new "Structure of intellect." In L. B. Resnick (Ed.) *The nature of intelligence*. Hillsdale, N.J.: Lawrence Erlbaum, 1976, 27-56.

Corts, D. B., Muldrow, T. W., & Outerbridge, A. N. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job success for customs inspectors* (TS-77-4). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, December 1977. (NTIS PB 280-260)

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 52, 281-302.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs*, 1951, No. 5.

French, J. W., Ekstrom, R. M., & Price, L. A. Manual for kit of reference tests for cognitive factors. Princeton, N.J.: Educational Testing Service, 1963.

Hunt, E., Lunneborg, C., & Lewis, J. What does it mean to be high verbal? *Cognitive Psychology*, 1975, 7, 194-227.

Munoz-Colberg, M., & Nester, M. A. *The measurement of inductive abilities according to the schemata of inductive logic* (TS-78-2). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, October 1978. (NTIS PB 293 306)

Northrop, L. C. *The definition and measurement of judgment* (TM-76-17). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, November 1976. (NTIS PB 268 942)

Northrop, L. C. *The definition and measurement of numerical ability* (TM-77-9). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, July 1977. (a) (NTIS PB 274 575)

Northrop, L. C. *The definition and measurement of reasoning* (TM-77-10). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, July 1977. (b) (NTIS PB 273 149)

Northrop, L. C. *The definition and measurement of verbal comprehension* (TM-77-11). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, July 1977. (c) (NTIS PB 274 572)

Northrop, L. C. *Documentation of the constructs used in Test 700 (Federal Service Entrance Examination)* (TM-79-12). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, August 1977. (NTIS PB 274 574)

Northrop, L. C. *Experimental item types to measure judgment* (TM-77-14). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, August 1977. (e) NTIS PB 274 573)

O'Leary, B. S. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of training success for social insurance claims examiners* (TS-77-5). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, September 1977. (NTIS PB 274 580)

O'Leary, B. S., & Trattner, M. H. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for internal revenue officers* (TS-77-6). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, September 1977. (NTIS PB 274 579)

Reck, M. *Criterion-related validity of PACE constructs: A review of the literature* (PS-77-4). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, September 1977. (NTIS PB 274 006)

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Trattner, M. H., Corts, D. B., van Rijn, P. P., & Outerbridge, A. N. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for claims authorizers in the social insurance claims examining occupation* (TS-77-3). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, September 1977. (NTIS PB 273 118)

U. S. Civil Service Commission. Tests and other applicant appraisal procedures. *Federal Personnel Manual System (FPM Supplement 271-2).* Washington, D.C.: Author, 1972.

U. S. Equal Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Justice, U.S. Department of Labor, U.S. Department of the Treasury. Uniform Guidelines on Employee Selection Procedures, *Federal Register,* 1978, 43(166), 38925-38309.

U.S. General Accounting Office. *Federal employment examinations: Do they achieve equal opportunity and merit principle goals?* (Report No. FPCD-79-46). Washington, D.C. Author, 1979.

Wing, H. *Normative data for the Professional and Administrative Career Examination (PACE): FY 1975* (TM-76-15). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, December 1976. (NTIS PB 286 102)

Wing, H. *PACE certification and selection in three area offices, Summer 1977* (TM-78-4). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, July 1978. (NTIS PB 286 102)

Wing, H. *Stability of characteristics of alternate forms of a test battery* (TM-77-7). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, August 1977. (NTIS PB 273 118)

PRDC PUBLICATIONS CONCERNING THE PACE

Corts, D. B. *Job Information test development for claims authorizers in the Social Security Administration* (TM 77-1). February 1977.

Corts, D. B., Muldrow, T. W., & Outerbridge, A. N. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job success for customs inspectors* (TS-77-4). December 1977. (NTIS PB 280 260)

Gavin, A. T., & Martin, C. G. *A procedure for estimating the unique contribution of a composite test: Uniqueness analysis of Test 500* (TM-76-8). (NTIS PB 261 701)

Martin, C. G. *A procedure for estimating the reliability of a weighted linear composite test: The reliability of Test 500, Series 110* (TM-75-4). July 1975.

Martin, C. G. *A procedure for estimating the reliability of a weighted linear composite test: The reliability of Test 500, Series 120* (TM 76-1). February 1976.

Martin, C. G. *A procedure for estimating the reliability of a weighted linear composite test: The reliability of Test 500, Series 130* (TM-76-2). February 1976.

Martin, C. G. *A procedure for estimating the reliability of a weighted linear composite test: The reliability of Test 500, Series 140* (TM-76-3). February 1976.

McKillip, R. H., Trattner, M. H., Corts, D. B., & Wing, H. *The Professional and Administrative Career Examination: Research and Development* (PRR-77-1). April 1977.

Munoz-Colberg, M. *A logical approach to the testing of deductive and inductive abilities* (TS-77-7). July 1977. (NTIS PB 274 569)

Munoz-Colberg, C. & Nester, M. A. *The measurement of inductive abilities according to the schemata of inductive logic* (TS-78-2). October 1978. (NTIS PB 293 306)

Munoz-Colberg, M., & Nester, M. A. *Utilization of taxonomic and other construction criteria in the written test portion of the Professional and Administrative Career Examination (PACE)* (PRR-79-5). August 1979. (In press)

Nester, M. A. *Time limits for competitive administration of the Professional and Administrative Career Examination* (TM-77-15). July 1977.

Northrop, L. C. *The definition and measurement of judgment* (TM-76-17). November 1976. (NTIS PB 268 942)

Northrop, L. C. *The definition and measurement of numerical ability* (TM-77-9). July 1977. (NTIS PB 274 575)

Northrop, L. C. *The definition and measurement of reasoning* (TM-77-10). July 1977. (NTIS PB 273 149)

Northrop, L. C. *The definition and measurement of verbal comprehension* (TM-77-11). July 1977. (NTIS PB 274 572)

Northrop, L. C. *Documentation of the constructs used in Test 700 (Federal Service Entrance Examination)* (TM-79-12). August 1977. (NTIS PB 274 574)

Northrop, L. C. *Experimental item types to measure judgment* (TM-77-14). August 1977. (NTIS PB 274 573)

O'Leary, B. S. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of training success for social insurance claims examiners* (TS-77-5). September 1977. (NTIS PB 274 580)

O'Leary, B. S., & Trattner, M. H. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for internal revenue officers* (TS-77-6). September 1977. (NTIS PB 274 579)

Reck, M. *Criterion-related validity of PACE constructs: A review of the literature* (PS-79-4). September, 1977. (NTIS PB 274 006)

Sapinkopf, R. C. *Statistical characteristics of the written test for the Professional and Administrative Career Examination (PACE) for visually handicapped applicants* (TM-78-1). January 1978. (NTIS PB 280 491)

Schmidt, F. L., & Hunter, J. E. *Development of a general solution to the problem of validity generalization* (TS-78-1). August 1978.

Trattner, M. H., Corts, D. B., van Rijn, P. P., & Outerbridge, A. N. *Research base for the written test portion of the Professional and Administrative Career Examination (PACE): Prediction of job performance for claims authorizers in the Social Insurance Claims Examining occupation* (TS-77-3). September 1977. (NTIS PB 273 118)

Wing, H. *Stability of characteristics of alternate forms of a test battery* (TM-77-7). August 1977. (NTIS PB 280 958)

Wing, H. *Normative data for The Professional and Administrative Career Examination (PACE): FY 1975* (TM-76-15). December 1976. (NTIS PB 286 102)

Wing, H. *PACE certification and selection in three area offices. Summer 1977* (TM-78-4). July 1978. (NTIS PB 286 102)

Wing, H. *Profiles of cognitive ability for different racial/ethnic and sex groups on a multiple abilities test battery* (TM-79-7). In press.

Wing, H. *Practice effects with traditional mental test items* (TM-79-9). In press.

DISCUSSION

Bond: How did you get the psychologists to make 1710 ratings?

McKillip: We locked them in a room. We gave them different offices in a nicer location, with comfortable surroundings. It took them about two weeks. They could do only about three occupations a day. They had to review the occupation, the classification standards, the qualification standards, the definitions of the abilities—all of that, before they started making the ratings.

Jackson: Did you have any evidence of the reliability of the ratings, the interjudge reliabilities?

McKillip: The ratings of the subject matter experts in the duty and time ratings averaged .91; the reliability of the ratings of the 31 KSAO's averaged .95 and as you might expect, the psychologists were the least reliable, the average being .72. That was because there were only six abilities to rate.

Gulliksen: You said you couldn't use information on race or that it was politically inadvisable—I don't know what. I mean, if you're going to avoid adverse impact, the only way to do it is to make use of information on race.

McKillip: This is a developing area. We were prohibited from collecting racial and ethnic data from applicants for a long period of time. Then with the acceptance of the Federal Executive Agency Guidelines in 1976, we were required to collect such data. We accomplished the task of determining how best to collect reliable data and are now doing so.

Gulliksen: So you can select?

McKillip: No. We can do studies of the selection procedure on racial and ethnic groups. But we cannot select on the basis of race.

Gulliksen: I mean, isn't this what they do in the Harvard approach, discussed favorably in the Bakke decision? They use test scores and go down to say, 500 or 300 or some score and label these people competent. Then since this results in three times as many applicants as they can admit, other criteria are used such as where do you come from, what's your race, can you play football and different things like that; having said first that you're competent.

Wing: The Catch-22 for the Federal Government is the second half of that Civil Service Reform Act principle which says that selection shall be determined solely on the basis of relative ability, knowledge and skills, after fair and open competition which assures that all receive equal opportunity.

The emphasis on "relative" has been interpreted pretty strictly as we go in rank order. We have another problem, a second Catch-22, if you will, in the Veteran

Preference Act. This is a completely different law from the Civil Service Reform Act. It provides for giving extra points to veterans, once an eligibility rating is established; and disabled veterans are placed at the top of the registers.

Gulliksen: But there is some sort of provision against impact also, isn't there?

Wing: If there is adverse impact then the validation requirements of the Uniform Guidelines apply.

Gulliksen: So you can have adverse impact if you satisfy some other requirements?

McKillip: Theoretically, yes; politically and socially, no.

Gulliksen: It sounds like a mess. They ought to change the rules.

Tenopyr: Have you done any other work on ratings of jobs by various people?

Wing: Only to the extent that we have used multiple criteria in our criterion-related validity studies. These criteria are correlated. The supervisory ratings criteria tend to produce the lowest validity coefficients.

Tenopyr: In my own research (with which I am very uncomfortable) encompassing subject matter experts; that is, supervisors rating abilities required in jobs, you get high reliability. Let's say that oral fluency is required of the telephone service representative; yet, you do a separate study and find that nobody is ever fired for lack of oral fluency.

Wing: That doesn't mean it isn't related to job performance

Tenopyr: The supervisors told us that was the main reason for turnover. But in other studies we have found, for example, in our financial area, that nobody did computations. You get tremendous response biases with respect to job stereotypes. High level people did not make computations, yet computational fluency tests were the best predictors for the job.

I am concerned about the possible biases that do creep in. I think it's an area for a great deal of further research.

McKillip: We had some pretty good results in the job analysis part of this study that are consistent and reasonable. The psychologists' ratings and the development of ability importance ratings were supported by the subject matter expert ratings of the 31 KSAO's. As far as the importance of the various KSAO's, we agreed with them.

Schoenfeldt: I want to ask what is the long-term future for the PACE.

McKillip: That's a good question. It is currently under very serious discussion, and who knows? There have been various proposals for changing the way the test is used: to use the test on a pass-fail basis with a low cutting point, and selecting or ranking people on some other basis, such as, evaluations of education or experience. These alternatives would be up to the individual agency evaluating applicants.

Question: Is it feasible in your organization to study the differential impact of certain item types?

Wing: It's intended. We have some evidence on the different test parts by racial and ethnic groups, and there is not much except between Asians and non-Asians. Asians, as distinct from others, tend to score higher on non-verbal than on verbal ability tests. There are racial/ethnic differences in overall level of performance, but the only differential is the verbal-non-verbal distinction between Orientals and others. This is also the only consistent finding in the literature.

Hunter: Real Orientals?

Wing: Persons who classify themselves as Asians, Pacific Islanders or from the Pacific Subcontinent.

Maslow: If administration of PACE is decentralized would agencies be permitted to borrow the basic examination and use it within the exam structure?

McKillip: Current proposals require administration of the PACE test. People who pass would receive an eligibility certificate or a hunting license. They could then apply at various agencies for jobs. The agencies would select from them in accordance with some procedure that they would establish and we would approve—"we" being the Office of Personnel Management.

Question: Can the agency give another test?

McKillip: That issue hasn't been raised. The answer will almost certainly be no.

Kraft: Agencies might use work sample tests, but I wouldn't know what work sample test might be appropriate to these occupations.

Wing: A simulation might be appropriate.

Question: Like hiring for six months.

Kraft: That's what the EEOC recommends.

Schoenfeldt: It seems to me that shifting to the use of the PACE test as a screening mechanism actually represents quite a different application. I'm not sure what the ramifications are, but, if you were setting up to do that back in the early Seventies, you probably would have approached it differently.

McKillip: We wouldn't have needed a three-hour, six-part test to do that. The effect would be the loss of many millions of dollars annually in reduced productivity of Federal employees.

Hunter: You are sure that wouldn't be billions rather than millions? In one validity study, I estimated the improvement in productivity resulting from use of the test to be $200,000,000 a year for the Philadelphia Police Department. The validities there were lower than the validities you have, so I think the more accurate figure would be well into the billions.

McKillip: A few years ago, we made several utility estimates, based on 1975 hires. We found that selecting randomly from the top half of the examinees as opposed to the top ten percent would result in a loss of $250 million per year. That would represent about 55 percent of the improved productivity obtained by selecting from the top ten percent.

Schoenfeldt: That would reduce adverse impact?

McKillip: There would still be a significant adverse impact, if you dropped the passing point down to the mean. There is something in the neighborhood of a 1.3 standard deviation difference in the average test scores of blacks and whites.

There might be a sufficient number passing if you could reach them, but they would be concentrated more at the bottom end of the register.

Schmidt: The cutting scores now being proposed are around one standard deviation below the mean?

Wing: Reducing the cutting score to 1 standard deviation below the mean would result in 33% of the blacks and 86% percent of the whites passing.

The racial and ethnic data we have now suggests we could maintain the current passing point, and still have a sufficient number of applicants eligible in each of the racial/ethnic groups.

If we were to develop a quota system which called for selecting 15% minority group members in their relative rank order, and selecting the others in the usual way, also in relative rank order, the loss in productivity would be something like four and a half percent as opposed to 55% when there is random selection from the top half.

Hunter: That's what we found also.

McKillip: The question is, is it worth about $25 million a year in productivity to have a representative work force? I think that if our political leaders would address that question, they would come to a very quick decision that it was well worth it. There still might be constitutional problems.

The pressure is to achieve a representative workforce. It seems to us you would be a lot better off by hiring the best of each group.

Question: Wouldn't you run into a Bakke-like case?

McKillip: We're in court anyway.

Gulliksen: My understanding of the Bakke case was that they had a separate quota for a given group, and that was illegal. The system that I described a moment ago is, as far as I know, not objectionable. That is, you don't have a quota, you just use various kinds of information.

McKillip: You just do what you have to do to reach your quota.

Gulliksen: But that's the problem, how you do it. You don't specify that you have a quota. In the Bakke case there was a quota, the university did not specify that the people that they selected in the minority group were competent.

McKillip: The University of California system, as I recall, was not geared toward the selection of racial or ethnic groups. Wasn't it geared toward disadvantagedness or something like that? But the way it worked in actual practice was selection on the basis of race.

Gulliksen: The point was they selected so many from a given group, regardless of the score. And the university system I described was to put a competency score down and say, Well, all these people are competent and they can graduate from the University. You don't take a given group and say, I'm going to select 20 people from that group regardless of how terribly bad they are.

Wing: It would be interesting to see how people who scored at the lower end of that qualified group performed in college, and what the dropout rate was.

Gulliksen: Yes.

Messick: In your predictive studies, I understand the six scores were combined, using the ability weights you obtained from the judgmental studies; that is, combining those with some weights that were obtained from job analysis and so on.

McKillip: Those were concurrent studies, and five abilities made up the final test battery. The correlations reported are between total test score and various criteria.

Horn: Those weights are different by occupation?

Wing: Yes, there are six occupational groups or clusters. Fifteen occupations were in Category A, 8 occupations were in Category B, and Category C, D and E and F each had one occupation.

Messick: Given the rationale and the intention of the tests, it would be informative to see what the pattern of correlations was of the five ability scores with each of the criteria.

Wing: The criterion-related studies do have those data reported in them. We calculated multiple correlations as well as correlations using the rational weighting system that we had devised, and unit weights. We found essentially no difference.

Hunter: I finally figured out why the utility figures you had were so low. You calculate the utility only of the new hires for one year. You have to figure either one of two ways. You have to figure on the basis of what you estimate to be the long-term effect once the test has basically selected most of the job pool, or, alternately, you have to take the expected tenure. If you want to know the impact of the new hires, you have to multiply the utility by the expected tenure of the employee. I would guess in your case that's probably 15 to 20 years, and that would then put your figure where I would guess it to be, which would be about $2.5 billion rather than $2.5 million per year.

McKillip: We wouldn't argue with that.

Schoenfeldt: I wanted to bring up a point. It seems to me that if you look at the chart that Bob Guion published a few years ago with the job content on the left and various measures across the top, it illustrates the inferential leap from the measure to the job content, probation and simple skills being the most job related. You would say that we could measure these by a content-oriented procedure. Then you move to a larger inferential leap with simple skill measures; such as, a math score on a math test as it relates to job performance. Finally you would move to things which would ordinarily require construct validity; such as, leadership, or some of the things you are measuring. It seems to me that if you can stick to measures that you can defend on a content basis, they may have adverse impact but it probably would be the smallest inferential leap. Job performance might also have adverse impact. When you go to criterion-related validation, you can show empirically that it works, and you may or may not want to address the constructs involved. This is also the case with content validation.

It seems to me what you have illustrated is the precariousness of trying to use construct validation. You are trying to make the largest inference and you are likely to have the largest adverse impact or at least an adverse impact.

McKillip: We don't think construct validity would have any larger adverse impact. In the area of the kinds of abilities that are pretty well documented in the literature—starting with the French kit and working from there—the inferential leap is more like a step over a crack in the sidewalk. We don't think it's that big. If you are dealing with constructs like leadership and other things where the literature is not so conclusive, then you have a good point. But here, it's a pretty easy step. People have to read and interpret complex written material and that takes verbal ability.

Guion: I want the record to show I have subsequently repudiated that chart.

ROBERT J. STERNBERG

THE CONSTRUCT VALIDITY OF APTITUDE TESTS: AN INFORMATION-PROCESSING ASSESSMENT

The closing months of 1979 provide an excellent time for yet another assessment of the construct validity of the available range of aptitude tests, because during this decade the construct validity of aptitude tests has undergone such intense scrutiny. These months provide a particularly appropriate time for an information-processing assessment of construct validity, because the information-processing view of construct validation has become so salient in the seventies. Experimental psychologists finally took to heart Cronbach's (1957) plea for a unification of the two disciplines of scientific psychology, and let come to pass Thurstone's (1947) prediction that "the rough factorial map of a new domain will enable us to proceed beyond the exploratory factorial stage to the more direct forms of psychological experimentation in the laboratory" (p. 56).

I will divide my assessment of the construct validity of aptitude tests into four parts. In the first, I will motivate the assessment by discussing why, to some at least, an information-processing assessment of the construct validity of aptitude tests has seemed like such a good idea. In the second part, I will present a brief historical overview of some of what has been done toward this goal in the last decade, and will assess some successes and failures of the attempts that have been made. A more detailed overview is provided by Carroll and Maxwell (1979). In the third part, I will briefly describe my own present componential view of aptitude. In the last part, I will explain what this view tells us about the construct validity of aptitude tests.

Psychometric Approaches To Construct Validation: Something's Missing

Psychometric approaches to construct validation of aptitude tests have generally sought to understand tests and test performance in terms of some kind of structural model, such as the factor-analytic model or the latent-trait model. Individual-differences variation in test performance is explained as deriving from one or more underlying structures, such as a general factor and specific factors, or a set of primary factors. These models have been criticized, rightly or wrongly, for a variety of sins, some of them venial, others, mortal. But for the present purposes, two aspects of the models are critical.

First, there is an element of circularity to these explanations that can become vicious when the explanations are used in construct validation. The aptitude tests under scrutiny were constructed on the basis of explicit or implicit psychometric theories of aptitude. These tests are supposed to provide construct validation of the aptitude theories, while at the same time the aptitude theories are supposed to provide construct validation of the aptitude tests. It's hard to have it both ways, if one's goal is an independent appraisal of the tests, of the theories, or, as is the case in construct validation, of both. The tests and theories may be conceptually compatible with each other, but neither can provide independent support for the other, because the interdependence between the two was established at the time the theories and tests were created. What would seem to be needed for construct validation of both the tests and the theories is some different kind of aptitude theory that is logically (although one would hope, not psychologically) independent of the psychometric tradition of theories and tests. Information-processing theories and tests seem to fill this bill.

Second, aptitude tests require and aptitude theories deal with dynamic performance. Solution of test items occurs over time, and the construct validity of both tests and theories will depend in part upon whether they can account for what happens over time. But although aptitude tests require dynamic performance, the indices of performance they provide are essentially static measures of terminal behavior on test items; and although aptitude theories must explain dynamic constructs, they generally explain these constructs in a static way, through structural models such as factor analysis. This does not mean that the theories are wrong or the tests inadequate. It does mean, though, that a complementary kind of theory and index of test performance are needed to deal with dynamic performance. The information-processing approach seems to provide what is needed. So let us turn to a consideration of just what we mean by information-processing theories and test indices.

Information-Processing Approaches to Construct Validation: An Historical Overview

What is an information-processing approach to construct validation? Indeed, what is an information-processing analysis? An information-processing analysis is an attempt to account for individuals' performance on one or more tasks in terms of the elementary processes used in task performance, the strategies for task performance into which these processes combine, and the representations for information upon which the processes and strategies act. An information-processing approach to construct validation seeks to examine the validity of psychometric tests and theories of aptitude through an analysis of the information-processing that occurs during test performance or that gives rise to a source of individual differences such as a factor. From an information-processing point of view, then, factors can be understood in information-processing terms (Carroll, 1976; Sternberg, 1977):

The two viewpoints are not inconsistent, but complementary.

Information-processing analysis of aptitude test performance during the seventies has arisen in part as a reaction to psychometric analysis of test performance (see Hunt, Frost, & Lunneborg, 1973; Sternberg, 1977). But if one goes back a decade further to the origins of information-processing analysis, one finds that it originated in large part as a reaction to stimulus-response analysis of task performance (see Miller, Galanter, & Pribram, 1960). At the time, many experimental psychologists had become dissatisfied with the inability of stimulus-response theories to account for complex forms of behavior. These theories seemed to account very well for the contingencies that brought about behavior, and for the responses that eventuated from these stimuli; but they were less well able to account for the processes that took place in the head and that served to link external stimuli and responses. Psychometric and stimulus-response accounts of complex behavior differ in many, many respects, but they have in common their seeming inability to provide dynamic accounts of information-processing. As its name implies, this is where the information-processing approach excels. What forms have information-processing analyses of aptitudes taken, and how have they functioned in the construct validation of aptitude tests?

The Aptitude-Treatment Interaction Approach

The aptitude-treatment interaction (ATI) approach dates back past the beginning of the seventies, at least to Cronbach's (1957) APA presidential address, where the ATI approach was suggested as one means of reuniting the two disciplines of scientific psychology. The approach provides construct validation by assuming that if an aptitude is psychologically real in some sense, and if a given test actually measures that aptitude, then the measured aptitude should interact with at least one instructional treatment in a meaningful way. For example, it might be proposed that high-reasoning subjects would do better to learn by discovery, whereas low-reasoning subjects would do better to learn by rules. If a given aptitude does not show a meaningful pattern of interaction with any instructional treatment, then this lack of empirical consequences for that aptitude (as measured) might lead one to question why the aptitude should be labeled an "aptitude" at all. Similarly, if two aptitudes never interact differentially with various instructional treatments, then the lack of differential empirical consequences for the two aptitudes might lead one to question whether they should be identified as separate aptitudes.

The history of ATI research has been exhaustively documented by Cronbach and Snow (1977). Hundreds of studies have failed to yield highly replicable findings, or to shed much light on the nature of aptitudes. Cronbach and Snow have suggested a number of technical inadequacies in past studies that are sufficient to account for the failures of so many ATI studies. But from our present vantage point, two interrelated elements seem to be key.

First, it has never been entirely clear how aptitude-treatment interaction research could be expected to yield a theory of aptitudes that would provide an independent basis for the construct validation of existing tests and theories. Rather, ATI methodology seems to provide a device for testing extant theories. It does not provide a method for decomposing test performance into elementary constituents of some kind, but rather a method for understanding overall test performance in an experimental, or, ideally, classroom context. Methodologies are needed for both functions, so this is certainly not a criticism of ATI methodology. It is merely a suggestion that ATI methodology does not provide the new, independent kinds of constructs needed for construct validation of psychometric tests and theories. Ideally, these new kinds of constructs could be employed in an ATI setting.

Second, if the pre-existing constructs that are used in ATI research are not useful for the purpose to which they are being put, then the research is likely to fail. Aptitude-treatment interaction experiments generally involve some instructional (or other experimental) treatment administered over time. The aptitude constructs used in the experiments should therefore be dynamic ones that are sensitive to the processes as well as to the products of performance. But the constructs used in ATI research have been, for the most part, psychometric ones that are not very sensitive to dynamic changes in information processing over the duration of the treatment. These psychometric constructs, no matter how useful they may be for other purposes, are probably not well suited for use in these experiments. A factorially-based measure of spatial ability, for example, will not be optimal if performance in response to a treatment involves spatial processes at some points in time, but, say, linguistic processes at other points in time (as is the case with certain types of syllogisms). If this argument is correct, ATI research should be more successful if dynamic, information-processing constructs are used instead of psychometric ones, assuming, of course, that the particular information-processing constructs used are theoretically appropriate and measured adequately. Two recent studies along these lines suggest that information-processing constructs can result in clear-cut aptitude-strategy interactions (Sternberg & Weil, 1980; Mathews, Hunt, & MacLeod, Note 1).

The Cognitive-Correlates Approach

What Pellegrino and Glaser (1979) have referred to as the "cognitive-correlates approach" can be traced back to an article by Hunt, Frost, and Lunneborg (1973). This line of research has been continued by Hunt and his associates throughout the course of the seventies (see Hunt, 1978; Hunt, Lunneborg, & Lewis, 1975). The basic idea is that performance on complex aptitude tests can be understood in terms of the information processing components that are used in the performance of relatively simple laboratory tasks studied by cognitive psychologists. For example, one such task requires subjects to compare whether two letters are both in the same case (upper- upper or lower-lower) or each in a different case (upper-lower or lower-upper). The subject would be timed, say, for how long it takes him or her to determine that "a A" represent different cases (cf. Posner & Mitchell, 1967). Another such task requires subjects to indicate whether a given target letter (or number) occurs in a short string of letters (or numbers). The target may be presented either before or after the string. The subject might be asked, say, to memorize the letter string "A-N-C-U-E." The subject would then be presented with a target letter, such as "U", and be asked to say as quickly as possible whether the letter "U" appears in the memory set. In both this task and the preceding one, time to respond is the major dependent variable. Using single tasks such as these to investigate verbal aptitude, Hunt et al. (1975) concluded that verbal aptitude can be understood largely in terms of the speed of access and retrieval of verbal information from short-term memory.

This kind of research has the potential for providing the kinds of constructs that are needed for construct validation of aptitude tests and theories. The cognitive theories under investigation are certainly independent of the psychometric tradition. The tasks were also developed independently of this tradition, however, and one might wonder whether they are the right tasks, and the processes that enter into them the right processes. I find it very difficult to believe, on merely intuitive grounds, that the simple tasks used in the cognitive psychologist's laboratory bear much relation to the complex tasks found on aptitude tests, or to the still more complex tasks required in everyday and in academic life. Although the laboratory cognitive tasks are quite nontrivial when examined in the context of certain cognitive theories, they appear to be rather trivial when viewed as a basis of performance for, say, comprehension of complex reading material. The experimental data Hunt and others have collected can be interpreted as supporting my reservations. Factor analyses of the aptitude scores and measures from the laboratory tasks reveal a major group factor comprising the various cognitive tasks (see Hunt et al., 1975). An interbattery factor analysis (Tucker, 1958) might be more illuminating, in that it would highlight only sources of variation shared across modalities (i.e. tasks versus tests). But I doubt the basic facts of the matter would change, since the aptitude tests tend to be rather highly intercorrelated, whereas the cognitive tasks are not very highly correlated with the aptitude tests, nor even, for the most part, with each other.

The Task-Analysis Approach

The task-analysis approach in experimental psychology dates back to Donders (1868); in the study of aptitudes, people date it back in different ways, but I see it as having originated with Johnson's (1960) method of serial analysis. In task analysis, the task of interest is decomposed into smaller constituents of information processing, usually, component processes. The basic idea is that the way to understand complex aptitudes is to take the test items that are used to measure these aptitudes, and to decompose them into their basic component processes. Different methods have been proposed for decomposing tasks, including my own method of "componential analysis" (Sternberg, 1977, 1978a, 1979a). Componential analysis was self-consciously created with construct validation in mind. "From a differential viewpoint, componential analysis may be viewed as a detailed algorithm for construct validation" (Sternberg, 1977, p. 65). I will not discuss each of the steps of the algorithm here, since they have been presented in some detail elsewhere (Sternberg, 1977). The major steps involve decomposing a complex aptitude item into a set of component processes, and relating these processes to global scores on a reference ability test.

Consider, for example, an aptitude item such as an analogy, and the construct of "analogical reasoning" that this item is supposed to measure. Factorial analysis would seek to understand performance on analogy items by relating this performance to performance on other kinds of test items (say, series completions and classifications), and by finding an underlying factor that serves as a latent source of individual differences on all of these items. Cognitive-correlates analysis would seek to understand performance on analogy items by relating it to performance on very simple laboratory tasks requiring manipulation of information in short-term memory, attention to perceptual properties, and so on. Task analysis would seek to understand it by decomposing analogy performance into a set of underlying component processes. I have suggested, for example, that an analogy such as "LAWYER is to DOCTOR as CLIENT is to (A. PATIENT, B. MEDICINE)" is solved by (a) *encoding* each analogy term, retrieving from long-term memory a set of semantic attributes relevant to each individual term; (b) *inferring* the relation between LAWYER and DOCTOR; (c) *mapping* the higher-order relation from the first half of the analogy (headed by LAWYER) to the second half of the analogy (headed by CLIENT); (d) applying from CLIENT to each of the two answer options, PATIENT and MEDICINE, the relation that was inferred from LAWYER to DOCTOR; (e) *justifying* one response as preferable to the other, even if nonideal; and (f) *responding*. These same components have been shown to be involved in the solution of other kinds of induction problems as well (Sternberg & Gardner, Note 2).

Task analysis, like cognitive-correlates analysis, provides an independent kind of theoretical construct for the construct validation of psychometric tests and theories. How well does the analysis fare? The question must be answered in two parts. First, we need to consider the internal validity of the analysis: How well do the proposed models of information processing in the tasks fit data collected from those tasks? Second, we need to consider the external validity of the analysis: How highly do task parameters correlate with external measures such as aptitude test scores, and are the patterns of correlation sensible?

From the standpoint of internal validation, the results of task analyses conducted during the past several years have been quite impressive. Models of task performance, although not accounting for all of the true variance in the response-time or response-choice data, have accounted for high proportions of the true variance in these data (e.g., Mulholland, Pellegrino, & Glaser, in press; Royer, 1971; Sternberg, 1977, 1980). From the standpoint of external validation, however, the results are more mixed. Major performance components of information-processing—such as the inference, mapping, and application components mentioned previously—have shown generally significant correlations with aptitude test scores (when they are applied to nontrivial analogies or other induction items), but the magnitudes have not been impressive. The response component, which one would not expect to be particularly highly correlated with aptitude test scores, has shown correlations that are both significant and of high magnitude. Moreover, this result has proven to be replicable in a number of different studies of different tasks by different investigators. It seems likely that it is not response itself, but something confounded in the response component, that is responsible for the high correlations. Response is estimated as a regression constant, meaning that the response parameter includes in its duration any operation whose duration is constant across all of the items being solved. Apparently, some of the most interesting sources of individual-differences variation have been lumped into a "constant."

Thus, although task analyses of the kinds that have been performed seem to be a reasonable start toward the construct validation of aptitude tests and theories, they are far from the finish. The realization that the task analyses I and others were performing just weren't extracting the most interesting sources of individual-differences variation led me to propose an augmented form of my own brand of task analysis, componential analysis, that extracts kinds of components of information processing that were not extracted in the earlier analyses. Some of the new methods my collaborators and I have proposed for extracting the various kinds of components are described in Sternberg (1978b) and in Sternberg (1979b). In the next section, I will briefly describe the theory underlying these new methods. This theory is proposed to apply to individuals across the whole range of levels of intelligence (Sternberg, 1979a, in press-a).

The Componential Theory of Intellectual Aptitudes: A Brief Overview

The proposed theory of intellectual aptitudes is component-based. Kinds of components can be classified in two different ways: by level of generality and by function (Sternberg, in press-b).

Levels of Generality

Components can be classified in terms of three levels of generality: General components are required for performance of all tasks within a given task universe; class components are required for performance of a proper subset of tasks (including at least two tasks) within a task universe; and specific components are required for the performance of single tasks within the task universe. I will ignore specific components, which seem not to carry their weight in a theory of intellectual abilities. Consider, for example, the components of analogical reasoning mentioned earlier. Suppose the universe of tasks is reasoning tasks of the kind found on many standardized tests of intellectual aptitudes. Encoding and response would seem to be general components, in that it is always necessary to encode the item information when it is presented, and to offer some kind of response after information processing on the item has been completed. Inference, mapping, application, and justification would seem to be class components. They are common to most inductive-reasoning tasks, i.e., tasks requiring individuals to make predictions of some kind about new information on the basis of old information, but are not found on most deductive-reasoning tasks, i.e., tasks requiring individuals to produce or select a single logically certain solution.

The three kinds of components may sound something like three kinds of factors: general, group, and specific factors. Components differ in at least three important ways from factors, however. First, their psychological reference is more clear than, and quite different from, that of factors. A component is a process with an experimentally determinable duration, probability of error, and effect upon response choice. The psychological reference of a factor, on the other hand, has never been very clear (see Sternberg, 1977). Just what does it mean for a factor, or anything else, to be a "latent source of individual differences"? Components are believed to be latent sources of individual differences, but their psychological properties can nevertheless be clearly specified. Second, components are identified in a very different way from factors. Components for an individual subject are identified by decomposing response-time, error, or response-choice data for that subject on a particular task. The outcome is independent of whatever other subjects or tasks happen to be used in a given experiment. The components contribute (additively or otherwise) to solution time, error rate, and response choice for a given type of task. Factors, on the other hand, are usually identified on the basis of individual differences in the data of multiple subjects. The outcome is dependent upon both who is in the sample and what particular tasks are being factor analyzed. Third, componential solutions are unique, whereas factorial solutions are usually subject to arbitrary rotation (unless they were identified through confirmatory maximum likelihood methods).

Function

Components perform (at least) five different kinds of functions. Each of these five functions can be crossed with the three levels of generality, yielding 15 different types (overall) of components. Five of these types of components—all specific ones—are not of much interest, leaving ten interesting kinds of components in the theory. Metacomponents are higher-order control processes that are used for planning how a problem should be solved, for making decisions regarding alternative courses of action during problem solving, and for monitoring solution processes. These metacomponents are sometimes referred to collectively as an "executive" or a "homunculus." Consider two examples of metacomponents, and of how they might function in an analogical-reasoning task. A first example is the selection of lower-order components. In the analogies task, an individual must select a set of component processes (such as the set presented earlier—encoding, inference, mapping, application, justification, response) that is sufficient for solution of the task. A second example is generation or selection of a strategy for combining these components. In the analogies task, an individual must decide how to sequence the set of components—for instance, should inference precede mapping, follow it, or perhaps occur simultaneously with it? Other metacomponents include selection of a representation for information, selection of a speed-accuracy tradeoff, and monitoring of one's solution processes.

Performance components are used in the execution of various strategies for task performance. The component processes named above for the analogies task are examples of performance components. Performance components may be viewed as actualizing the plans laid down by the metacomponents. For example, once the decision has been made that the relation between the first two terms of an analogy must be inferred, the inference component actually performs the job of inferring this relation.

Acquisition components are skills involved in learning new information; retention components are skills involved in retrieving information that has been previously acquired; transfer components are skills involved in carrying over retained information from one situational context to another. New information is always presented in some kind of a context, no matter how impoverished. In terms of situational contexts, acquisition components represent particular skills involved in utilizing context cues to learn new information; retention components represent particular skills involved in retrieving the information that was once learned; and transfer components represent particular skills involved in relating old contexts to new ones. Consider, for example, one aspect of multiple contexts, their variability across multiple presentations of a single piece of information, such as the appearance of an unfamiliar word. Some things that make an individual more "highly verbal," according to the present theory, are, first, his or her greater ability to use this variability of context to acquire the meaning of the word; second, the individual's greater ability to retrieve one or more contexts when the individual needs to retrieve the meaning of the word; and third, the individual's greater ability to carry over a context, which has now become an "old" context, to a new context to understand the meaning of the word when it is presented in an entirely new context.

Interrelations Among Kinds of Components

Kinds of components are interrelated in various ways. Consider first how components of different levels of generality are related, and then how components serving different functions are related.

General components are related across tasks by virtue of the fact that a general component occurs, by definition, in the performance of each task in the task domain under consideration. At the other extreme, specific components are each unique to individual tasks. Thus, most of what will cause some tasks to be more or less closely related to other tasks is the class components they do or do not share. Classes of tasks such as inductive-reasoning tasks, deductive-reasoning tasks, spatial visualization tasks, etc., tend to share many class components within the members of each class, but fewer class components across classes of tasks. Some classes of tasks will be more closely related than others by virtue of their sharing more class components with each other. (A more detailed description of just how the tasks are interrelated via class components can be found in Sternberg, 1979a, and in Sternberg, in press-b).

The various functional kinds of components are related to each other through the metacomponents. In the proposed system, metacomponents can directly activate, provide information to, and receive feedback from each other kind of component. A given metacomponent can also communicate with and activate other metacomponents. The other kinds of components can activate and communicate with each other only indirectly through the metacomponents, however. Thus all control and information in the system pass directly to and from the metacomponents. The metacomponents act as a filter in the relations of the other kinds of components to each other (see Sternberg, in press-b). For example, acquisition of information affects retention of information, but only via the link of acquisition and retention components to metacomponents, not via any direct link of acquisition and retention components to each other.

The metacomponents are able to process only a limited amount of information at a given time. In a difficult task, and especially in a *new* and difficult one, the amount of information being fed back to the metacomponents may exceed their capacity to act upon this information. In this case, the metacomponents become overloaded, and valuable information that cannot be processed may simply be wasted. The total information-handling capacity of the metacomponents of a given system will thus be an important limiting aspect of the system. Similarly, capacity to allocate attentional resources so as to minimize the probability of bottlenecks will be part of what determines the effective capacity of the system (see also Hunt, Note 3).

Construct Validation of Aptitude Tests via the Componential Theory

The componential theory of intellectual aptitudes sketched in the previous section can provide a basis for the construct validation of aptitude tests and the factorial theories upon which these tests have been based. I propose that what seems like conflicts between theories disappears when these theories (and the tests they have generated) are viewed through the componential lens.

How is it possible for more than one psychometric theory of intelligence to be correct? It is possible because these theories differ from each other primarily in terms of the way a given factorial solution is rotated. The choice of a rotation is mathematically arbitrary, although different rotations of a given factorial solution seem to highlight different psychological elements. I have claimed previously that the choice of a rotation is a matter of convenience (Sternberg, 1977). It depends upon what distribution of components one wishes to highlight. Consider how various psychometric theories can be interpreted in terms of the particular set of components they render salient.

Spearman's Two-Factor Theory

According to Spearman's (1927) two-factor theory and tests derived from it, intelligence can be understood in terms of two kinds of factors: a general factor whose influence permeates all tests of intellectual aptitude, and specific factors that are each found only on single aptitude tests or tasks. Evidence in support of this theory tends to be obtained in unrotated factorial solutions. According to the componential theory, individual differences in the general factor are attributable to individual differences in the effectiveness with which general components are performed. In other words, the general factor comprises a set of general components that is common to a wide variety of intellectual tasks. Specific factors comprise specific components.

As it happens, the metacomponents have a much higher proportion of general components among them than do any of the other kinds of components, presumably because the executive routines needed to plan, monitor, and possibly replan performance are highly overlapping across tasks of a widely differing nature. Thus, individual differences in metacomponential functioning will be primarily responsible for the appearance of individual differences of a general nature. Metacomponents are not solely responsible for the appearance of a general factor, however. Most behavior, and probably all of the behavior exhibited on intelligence tests, is learned. Thus, there are certain to be acquisition components, as well as retention and transfer components, whose past influence will have a present effect upon individual differences in general ability. Finally, certain performance components, such as encoding and response, may be general to a wide variety of tasks, and thus also have an effect upon individual differences in the general factor.

Thurstone's Theory of Primary Mental Abilities

According to Thurstone's (1938) theory of primary mental abilities and the tests derived from it, intelligence can be viewed as comprising a small set of multiple factors, or primary mental abilities: verbal comprehension, number, spatial visualization, word fluency, perceptual speed, reasoning, and memory. A multiple factor solution such as this one tends to appear when factorial solutions are rotated to "simple structure." Simple structure solutions, like unrotated solutions, seem to have a special appeal to psychometricians, and I believe there is a psychological basis for this appeal. Whereas the unrotated solution seems to provide the best overall measure of individual differences in general components, a simple structure solution seems to provide the best overall measure of individual differences in class components. There seems to be minimal overlap in class components across factors when factors are rotated in this way. Although overlap among class components is minimized, one would nevertheless expect some correlation between pairs of primary mental abilities, due to their overlap in general components. Indeed, when simple-structure solutions are factored, they tend to yield a second-order general factor, which I believe captures the shared variation across primary mental abilities due to general components.

Consider, for example, the inductive-reasoning factor that appears in Thurstone's theory. I have claimed that a relatively small set of class performance components—inference, mapping, application, and justification—appear in a wide array of inductive reasoning tasks (Sternberg, 1979a). Studies of spatial ability suggest that a rather small set of class performance components can probably account for performance on a variety of spatial tasks (see, e.g., Shepard & Metzler, 1971). And our present research on verbal comprehension suggests the possibility that a relatively small set of class acquisition, retention, and transfer components may account for much of the individual-differences variation in verbal comprehension tasks.

Cattell and Horn's Theory of Fluid and Crystallized Ability

Cattell and Horn (see, e.g., Cattell, 1971; Horn, 1968) have proposed that general ability can be divided into two subabilities, fluid ability and crystallized ability. Fluid ability is best measured by tests of abstract reasoning, such as visual analogies, classifications, and series problems. Crystallized ability is best measured by tests of vocabulary, reading comprehension, and general information. A very similar theory has been proposed by Vernon (1971), whose major group factors of practical-mechanical ability and verbal-educational ability seem to correspond closely, if not exactly, to Cattell and Horn's factors of fluid and crystallized ability. Factors such as these seem to result from hierarchical forms of factor analysis. On the componential view, crystallized ability tests seem best to separate the products of acquisition, retention, and transfer components, whereas fluid ability tests seem best to separate the execution of performance components. I believe that the measurement of crystallized ability in typical testing situations differs in a key way from the measurement of fluid ability. Whereas the measurement of crystallized ability involves *accumulated products* of past executions of components of acquisition, retention, and transfer, measurement of fluid ability involves *current execution* of components of performance.

The difference in the way the two sets of components are measured may explain why measured crystallized ability tends to increase indefinitely with increasing age (short of senility), whereas fluid ability starts to decline in early to middle adulthood. Products of performance are already established and unlikely to show deterioration except through the effects of senility. Current execution of performance components or of any other kinds of components is likely to deteriorate with age, however.

Guilford's Structure-of-Intellect Theory

Guilford's (1967) structure-of-intellect theory claims that intelligence can be viewed as comprising 120 distinct intellectual aptitudes, each represented by an independent factor. Support for this theory can be obtained by Procrustean rotation of factorial solutions toward the prespecified theory. Horn and Knapp (1973) have cast a shadow over the psychometric, and hence psychological, acceptability of this theory. Nevertheless, I believe that there probably is a psychological basis to at least some aspects of this theory, and that these aspects of the theory can be interpreted in componential terms.

A given component must act upon a particular form of representation for information, and upon a particular type of information. The representation, for example, might be spatial or linguistic; the type of information (content) might be an abstract geometric design, a picture, a symbol, a word, etc. Forms of representation and contents, like components, can serve as sources of individual differences. A given individual might be quite competent when applying a particular component to one kind of content, but not when applying it to another. Representation, content, and process have been more or less c
founded in most factorial theories, probably becau
tain components tend more often to operate up
kinds of representations and contents, and c
nents tend more often to operate upon d

representations and contents. This confounding serves a practical purpose of keeping to a manageable number the factors appearing in a given theory or test. But it does obscure the probably partially separable effects of process, representation, and content. Guilford's theory self-consciously tries to provide some separation, at least between process and content. I doubt the product dimension has much validity, other than through the fact that different kinds of products probably involve slightly different mixes of components. On the one hand, the theory points out the potential separability of process and content. On the other hand, it does so at the expense of manageability. Moreover, it seems highly unlikely that the 120 factors are independent, as they will show overlaps, at minimum, in shared metacomponents.

Conclusions

To conclude, the componential theory can provide at least a tentative and sketchy account of how different forms of factor analysis and rotation can support different factorial theories of intelligence. On this view, each so-called theory can be viewed as a special case of a single theory. Each special case highlights different constellations of individual differences, and which constellation of individual differences is important will depend upon the purpose the theory or test is intended to serve. In an important sense, then, most of the theories are "correct." The componential theory provides a complementary perspective on the various differential theories.

Tests of intellectual aptitude have numerous failings, but when all is said and done, there are very few nontrivial predictions of any kind that we can make in life that are a great deal more accurate than those made by aptitude tests. Many people, in fact, are willing to acknowledge that aptitude tests have been the most successful technological innovation that has arisen from psychological theory and research. If this is merely an unhappy commentary on what psychology can produce, so be it. But I think this is not the case. During the seventies, experimental psychologists have embarked upon an intensive effort to examine the construct validity of tests and theories of intellectual aptitude. My retrospective view of the outcomes of this research is that it has supported rather than refuted their construct validity. Indeed, we knew the tests were working rather well; now we have, perhaps, a better idea of why. This is not to say that the tests are perfect, although I suspect that many of the imperfections that people complain about most bitterly reside not in the tests, but in people's sometimes unrealistic notions about what the tests can do. There is much more to intelligence than is measured by the relatively narrow focus of current aptitude tests, and there is much more to academic and life achievement than could be measured by any kind of intelligence test. If the validity coefficients of aptitude tests have not risen appreciably in the past thirty or forty years, it may be because we have reached the limits of what measurement of a narrow set of intellectual aptitudes can predict.

It is quite possible that eventually we may have tests that measure some of the kinds of components I have discussed above. Such tests would serve an important theoretical purpose that is not being served by current tests. They would be measuring performance at a more basic level than is possible through existing test batteries. The question of practical value will not really be answerable until the tests are in a usable form, although I suspect such tests would supplement rather than replace the descendants of our present kinds of tests. In the meantime, I believe we would do well to continue the basic information-processing research we have been doing, since it seems to be bringing us closer to an understanding of what intelligence is, and of what intelligence tests measure.

Notes

1. Mathews N. M., Hunt, E., & Macleod, C. *Strategy choice and strategy training in sentence-picture verification.* Manuscript submitted for publication, 1979.
2. Sternberg, R. J., & Gardner, M. K. *Unities in inductive reasoning* (NR 150-412 ONR Technical Report No. 18) New Haven: Department of Psychology, Yale University, 1979.
3. Hunt, E. B. *Intelligence as an information processing concept.* (NR 154-398 ONR Technical Report) Seattle: Department of Psychology, University of Washington, 1979.

References

Carroll, J. B. Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), *The nature of intelligence.* Hillsdale, N.J.: Erlbaum, 1976.

Carroll, J. B., & Maxwell, S. E. Individual differences in cognitive abilities. *Annual Review of Psychology,* 1979, *30,* 603-640.

Cattell, R. B. *Abilities: Their structure, growth, and action.* Boston: Houghton-Mifflin, 1971.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist,* 1957, *12,* 671-684.

Cronbach, L. J., & Snow, R. E. *Aptitudes and instructional methods.* New York: Irvington, 1977.

Donders, F. C. Over de snelheid van psychische processen. Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868-1869, Tweede reeks, *II,* 92-120.

Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.

Horn, J. L. Organization of abilities and the development of intelligence. *Psychological Review*, 1968, *75*, 242-259.

Horn, J. L., & Knapp, J. R. On the subjective character of the empirical base of Guilford's structure-of-intellect model. *Psychological Bulletin*, 1973, *80*, 33-43.

Hunt, E. B. Mechanics of verbal ability. *Psychological Review*, 1978, *85*, 109-130.

Hunt, E. B., Frost, N., & Lunneborg, C. E. Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 7). New York: Academic Press, 1973.

Hunt, E. B., Lunneborg, C. E., & Lewis, J. What does it mean to be high verbal? *Cognitive Psychology*, 1975, *7*, 194-227.

Johnson, D. M. Serial analysis of thinking. In *Annals of the New York Academy of Sciences* (Vol. 91). New York: New York Academy of Sciences, 1960.

Miller, G. A., Galanter, E., & Pribram, K. H. *Plans and the structure of behavior*. New York: Holt, Rinehart, & Winston, 1960.

Mulholland, T., Pellegrino, J. W., & Glaser, R. Components of geometric analogy solution. *Cognitive Psychology*, in press.

Pellegrino, J. W., & Glaser, R. Cognitive correlates and components in the analysis of individual differences. In R. J. Sternberg & D. K. Detterman (Eds.), *Human intelligence: Perspectives on its theory and measurement*. Norwood, N. J.: Ablex, 1979.

Posner, M., & Mitchell, R. Chronometric analysis of classification. *Psychological Review*, 1967, *74*, 392-409.

Royer, F. L. Information processing of visual figures in the digit symbol substitution task. *Journal of Experimental Psychology*, 1971, *87*, 335-342.

Shepard, R. N., & Metzler, J. Mental rotation of three-dimensional objects. *Science*, 1971, *171*, 701-703.

Spearman, C. *The abilities of man*. New York: Macmillan, 1927.

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, N.J.: Erlbaum, 1977.

Sternberg, R. J. Componential investigations of human intelligence. In A. Lesgold, J. Pellegrino, S. Fokkema, & R. Glaser (Eds.), *Cognitive Psychology and instruction*. New York: Plenum, 1978. (a)

Sternberg, R. J. Isolating the components of intelligence. *Intelligence*, 1978, *2*, 117-128. (b)

Sternberg, R. J. The nature of mental abilities. *American Psychologist*, 1979, *34*, 214-230. (a)

Sternberg, R. J. A review of "Six authors in search of a character: A play about intelligence tests in the year 2000." In R. J. Sternberg & D. K. Detterman (Eds.), *Human intelligence: Perspectives on its theory and measurement*. Norwood, N. J.: Ablex, 1979. (b)

Sternberg, R. J. Cognitive-behavioral approaches to the training of intelligence in the retarded. *Journal of Special Education*, in press. (a)

Sternberg, R. J. Sketch of a componential subtheory of human intelligence. Behavioral and Brain Sciences, in press. (b)

Sternberg, R. J. Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, 1980, *109*, 119-159.

Sternberg, R. J., & Weil, E. M. An aptitude-strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology*, 1980, *72*, 226-234.

Thurstone, L. L. *Primary mental abilities*. Chicago: University of Chicago Press, 1938.

Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.

Tucker, L. R. An inter-battery method of factor analysis. *Psychometrika*, 1958, *23*, 111-136.

Vernon, P. E. *The structure of human abilities*. London: Methuen, 1971.

Footnote

Preparation of this report was supported by Contract N0001478C0025 from the Office of Naval Research. I am grateful to Janet Powell and Sandra Scarr for comments on an earlier version of this manuscript, and to my graduate-student collaborators in research, who have helped shape my views on the construct validation of aptitude tests.

DISCUSSION

Horn: You raised a lot of questions with me, but I would like to get started on this briefly, at least, by going back to Hulin's remarks yesterday. It seems to me that they were giving short shrift to the facet view that could be attributed to Guilford and certainly to Jack Carroll and some others. They were arguing, in effect, that you could just take any collection of abilities and put them together somehow, and if you took enough of them you would get a measure of general intelligence. And he was using the facet theory notion to satisfy that. Well, all that is by way of a plan for writing test items or something of that sort. It seemed to me the reason why that was an unfair characterization of, at least, Jack Carroll's use of that theory—and I think perhaps Guilford's too—is that I believe that the facet theory approach is also a componential approach in much the same sense as your approach.

I think what people are trying to do is a logical analysis, the way that information can, for example, get into the individual, and the way the individual can then operate on this information. And that is an attempt to build a theory about how individuals think, which is what construct validity comes down to being, when you are thinking in terms of tasks that psychologists have invented as opposed to the criteria that tasks would try to predict.

Well, with that kind of introduction, it seems to me your approach here is really not recognizing several of the facets that would be, I think, observed by psychometricians to be important. For example, the distinction between auditory and visual input is not at all handled by the notion of your five classes of components, or at least so it seems to me. It seems you would have everything going on somewhere centrally, and it wouldn't allow for the organization that could very well be coming through, say, the vision modality. How do you deal with that?

Sternberg: Let me just say two things: One is that I think Lloyd Humphreys really has been very concerned with facets. In 1962, he had an American Psychologist paper where he proposed to view Guttman's facet theory hierarchically. Although it may not have come out in the present paper, I think Lloyd would be sympathetic to the kinds of things you were saying initially.

Second, I don't mean to ignore facets. I agree that mode of presentation can, under certain circumstances, be important. In the literature on recall and recognition, sometimes mode of presentation counts and sometimes it doesn't. The problem one comes into is that when you start crossing all possible facets, you end up with an exceedingly large number of cells. This is a problem we all have to deal with.

My way of dealing with it is to say that there are basic kinds of components that have different functions and different levels of generality, and that when they are crossed, one ends up with ten kinds of interesting components and five kinds of boring ones.

One must look at the operation of these components on different kinds of representations and one must also look at how the information entered the system. Auditory or visual processes can then be sources of individual differences in information entry.

So I guess my tentative solution would be to have a more manageable number of components that can operate upon different kinds of representations and in different modes, and in that way retain some degree of parsimony while at the same time acknowledging multiple potential sources of individual differences other than components.

Horn: Just to follow-up, one reason why facet theory fails, or has seemed to fail, is not so much that there are so many facets—it's just like Guilford got to a logical kick, so to speak, when you start thinking facets, you can think forever. But until you say, "What's the processing interpretation, what would it mean in terms of almost any intellectual problem that would come to people?"—unless you come back that way, I think you are in trouble. That was the point of Humphreys' criticism of that viewpoint as it came out. And you are right, when he looked at Guttman's attempt to use facet theory, then it got back much closer to a process analysis of behavior.

Sternberg: I guess you have to take them into account without allowing your theory to explode exponentially, and that's what I tried to say.

Messick: Concerning your analysis of Guilford's 120 factors: Guilford would say he was concerned with a kind of process analysis since he would claim his operations are processes. He happened to cross them with the kind of information he thought existed in a structured environment. And in his 1967 book, he gives what he called "operational models," combining the five operations which, incidentally, include an operation for acquisition, encoding, and comprehension, which he called "cognition," and an operation for retention, which he called "memory;" two kinds of operations with related transfer, because he thought new information could be used in two ways: by convergence and by divergence. His models were taking those processes, combining them to account for still more complex and higher-order sequential tasks—such as problem-solving or creative production, which seems to be a level above the kind of componential analysis you are talking about.

Sternberg: There are so many things in there I want to respond to, I should have taken notes.

The first thing I want to say is that I agree with your assessment of Guilford. I think right now there is a certain tendency to throw out the baby with the bath because Guilford did one thing that might have been unfortunate—to have so many factors. People forget he was one of the primary process theorists among the factor analysts. I couldn't agree more with that. I think that was an important contribution. As I said, I think a lot of what I said dovetails with different kinds of factor theories.

A second thing that also relates to a point made in Hulin's talk yesterday is that I still believe there is a problem with identifying processes via factor analysis, and it comes out most clearly when he said you can't use small numbers of kinds of items, because then you won't be able to extract sources of variance.

One thing I believe is that in a methodology that is valid for extracting processes, one can take a single kind of item and extract the processes—the acquisition components, performance components, metacomponents, and so on—without having other kinds of items that affect what comes out. In factor analysis, what comes out depends on the context surrounding what you put in. That doesn't happen in componential analysis. And I think that's important.

People who are good on one process tend to be good on another. But, as we know, correlations across persons and correlations across task manipulations can be zero. They are not necessarily related, and you may find two distinct processes that in fact have high correlations.

The last part of what you said—and I think in some ways, the most interesting—concerns the application of componential analysis beyond very simple kinds of test items. Right now, we are doing research intended to show such applicability. We are working on complex types of causal inference of the kind an epidemiologist would use, or a stock analyst would do for a market analysis. I think the problems are fairly realistic in terms of what people in those fields have to do.

In our work on acquisition, transfer, and retention, we are using articles from newspapers, textbooks, and magazines with very low-frequency words embedded in them that are quite like materials children encounter in schools, and like materials we encounter when we read newspapers. In short, I hope that componential analysis can be carried out on higher-order types of tasks.

Hunter: I would like to make a criticism and that's the criticism that by constantly coming up from the lowest level and trying to build up, what happens is you miss a great deal. For example, I think your treatment of metacomponents adds virtually nothing to the standard treatment of factors as was done in factor analysis. It adds nothing, for example, to what Guilford did.

I thought your analysis of fluid and crystallized ability adds nothing to what Cattell has already said. And I think the reason is that, if you focus on small tasks, what happens is you would have oversimplification. For example, you seem to claim that acquisition is a simple process and is the same for everybody, and this is simply false, if you start by looking at complex tasks. And for any of you who know the concept-identification literature, if you contrast the treatment of concept identification—Bruner, Goodnow, Austin, Restle—the people in the Restle tradition who started from low-level concepts were never able to handle the kinds of complex tasks that Bruner, Goodnow, and Austin analyzed to begin with.

Let me give you an example in terms of retention. I have done a number of studies with graduate students on different kinds of processes and memory that are involved in complex tasks. For example, we have people learn a complex spy network and then begin by giving tasks from certain spies to other spies, or ask what would happen if the spy network were broken. And what you find is that retention is not a simple matter which is characterized by a quantitative parameter as would be assumed from actually anywhere in the classical background or from your work as well, but, rather, there are qualitative differences.

We found three completely different types of memory structures that people build. Some people try to learn the network using a paired associates method. What they try to do is to erect a structure in which each spy is related to whomever he can communicate with. Some people memorized lists. They simply took information about who communicated to whom, and mastered the lists. And when it came time to operate—and, I might add, these were the fast-test people—there were some people who tried to construct what turned out to be a motor or perceptual structure for the spies. And using eye movements, they more or less put this into a spatial configuration. They were radically different in terms of reaction time; while they did differentiate between the people, they did not have quantitative meaning. Reaction times meant the operation of what were actually different processes, not just differences in the same processes.

Gulliksen: Could the people who utilize any one of these three methods you have described be instructed that they shouldn't do that, and should use another method?

Hunter: We didn't actually try that. The research was exploratory and there were also great differences in numbers. The paired-associates processors constituted something like 75 percent of the data, so we had only a small number of the others. We did talk about what strategies they used in other problems. For example, the list processors typically used list processing. If we asked them, for example, how they operate in terms of schoolwork, they did lots of memorization. They found it convenient to set things up on lists, memorize them, and assess them because they were extremely fast at processing them. And the structural people did make a habit of trying to come up with ways of organizing material they had to learn in terms of pictures or in terms of other things.

Gulliksen: But you don't know if people could switch?

Hunter: No. My guess is they probably could but, on the other hand, there are those qualitative differences in preference.

Williams: Can Bob respond?

Sternberg: There are two aspects of what you said I could try to respond to. In terms of the first part, dealing with what, if anything, componential investigations add to factor analytic ones, I guess what I would hope they would add is a reasonably direct way of uncovering what the higher-order control processes or metacomponents are in people performing tasks factor analysts have studied. The methodology also isolates other kinds of components. If that can be done by factor analysis, then I would be the first to say factor analysis is performing a really valuable function, and I will start using it more myself. I haven't seen how Guilford or anybody else has done that using factor analysis, and I'm sure this isn't the place for you to show me. But if that can be shown to me and to others, I'll hop on the factor analytic bandwagon. Maybe I am missing something, but I guess I do think componential analysis provides a method for isolating processes that factor analysis doesn't provide.

The second part of your statement, I think, reflects a misunderstanding as to what componential analysis or forms of information-processing analysis can do. I'm not saying that there aren't major individual differences in components, strategies, representations, and so on. In fact, in my book and other publications, I have made it fairly clear, I think, that I view those as major sources of individual differences. And, moreover, a componential type of analysis can uncover those types of differences.

Let me give an example of what I am talking about. We did a study in which we had people solve problems such as: "Lillian is taller than Mary. Mary is taller than John. Who is the tallest?" We trained some of them to use an algorithmic strategy involving minimal reasoning. It's based on a flow chart that's quite simple to follow. And we had an untrained group and a visually trained group. What we were interested in was whether specific strategies could be taught, and also whether some strategies were better for people with certain aptitude patterns; in other words, does the optimal strategy for a person vary as a function of his or her ability pattern?

The next thing we needed to know is what did the people actually do? Well, how do you find out what they did? What most people do is what they were trained to do, or what they say they do. The two problems are, one, not everyone does what they are trained to do; and, two, on problems like these and many other kinds of problems used on tests, people don't know what they did. Either they can't say or what they say is vague and one can't get much out of it.

Incidentally, it turned out we originally got no aptitude strategy interactions. There was no differential pattern of correlations for different groups. What we then did was to look at what each person really did by modeling each individual subject's data. We took each subject in each group, fit a spatial model, a linguistic model, a spatial-linguistic mixed model, and the algorithmic model; and it turned out that although most people did what they were trained to do, a fairly substantial minority did not. We then reanalyzed the data and what we found was that among those using the spatial strategy, there was a very high correlation with spatial ability but not with linguistic ability. Among those using the linguistic strategy, there was a high correlation with linguistic ability but not with spatial ability. Among those who used the mixed model, there was a high correlation with both kinds of abilities. And among those who used the algorithmic strategy, there was a moderate correlation with both abilities. It was only by analyzing each person's data componentially that we were able to find this sort of aptitude-strategy interaction. I would say that componential methodology can serve as a useful basis for training, telling us some of the kinds of components that we should be training. So I don't think we disagree.

Messick: Did you take that finding to imply spatial ability was engaged by the spatial subjects and linguistic ability was engaged by the linguistic subjects? Were the abilities being engaged as processes in the test?

Sternberg: I distinguish between representations and processes. In other words, if people are really using a spatial representation for information here, imaging Lillian, Mary and John, then the chances are that their representation would give rise to a pattern of individual differences highly consistent with that obtained, and similarly for a linguistic representation. But, yes, I take the evidence as indicative of different representations, and different processes are needed to act upon these various representations.

DOUGLAS N. JACKSON

CONSTRUCT VALIDITY AND PERSONALITY ASSESSMENT

In this presentation my intent is to discuss the impact on personality assessment of the revolution in conceptions of testing heralded by construct validation. I will not seek to explicate the basic philosophical and psychological issues: that has been accomplished very well by others (e.g., Braithwaite, 1953; Cook & Campbell, 1976; Cronbach, 1970; Cronbach & Meehl, 1955; Dunnette & Borman, 1979; Hempel, 1952, 1965; Loevinger, 1957; Margeneau, 1950; Messick, 1979; Tenopyr, 1977; Wiggins, 1973). What I would like to emphasize are some of its implications for personality scale construction and for the prediction of behavior. The impact of this approach on my own thinking about assessment has been profound, both in the areas of test construction and in person perception. My message is that through a judicious combination of psychological analysis of dispositional variables and psychometric and multivariate procedures, progress in personality assessment is possible. In presenting this message I shall seek to marshall evidence that the use of implicit conceptions of personality trait and behavioral relationships, especially when represented by the combined wisdom of a number of judges, has potential for repaying handsome dividends in scale construction, in behavioral predictions, and in providing perspectives on the personological requirements of criterion performance.

Although there has been a resurgence in the application of nonintellective measures to personnel selection (e.g., R. C. Edwards, 1977; Howell, 1966; Landy, 1976), my purpose is not to argue for or against the utility of such measures in particular situations. Indeed, probably the majority of these studies have proceeded in the classical traditions of psychotechnology, employing empirical rather than hypothesis-testing approaches to measure development and selection and to the evaluation of criteria. A strictly empirical approach is limited in its generalizability and in its bearing on psychological theory (Loevinger, 1957). Such studies do serve to illustrate, however, that some portion of the substantial criterion variance not accounted for by intellective measures may otherwise be predictable. It does not follow that the dearth of positive evidence for the empirical validity of nonintellective measures identified in past reviews necessarily extends to the present and the future.

Construct Validity and Test Construction

In no area of psychological measurement has the influence of the construct validity approach been greater than in the area of personality. The traditional wisdom of test construction involves the (sometimes implicit) specification of a content domain, development of items to sample representatively processes related to content, and some means of empirical item selection, usually based upon measures of association between each item and a total or criterion score. A construct approach to personality scale construction involves a considerably more elaborate effort to link personality items differentially to constructs, one in which definitional and judgmental steps play a crucial role (Jackson, 1971; Wiggins, 1973).

The definition of a trait or construct in relatively precise, even if abstract terms, is an important first step. (Of course, it is necessary to choose constructs permitting specific, explicit definitions. Unfortunately, some of the most popular personality trait measures show little evidence of such a choice.) If definitions are adequate, my experience is that items can be prepared with relative dispatch, even by persons with modest formal psychological training. In a sense, educing and specifying valid relationships between a construct definition and an explicit behavioral exemplar, as for example, that contained in an item, is an extension of the construct definition. Defining a trait like dominance by successive approximating in terms of a large set of specific items is thus analogous to Hempel's (1952) suggestion that defining an empirical construct like temperature might involve outlining its known effects upon different physical materials and events. The advantage of an abstract definition is that it may suggest not only known manifestations of the construct, but hypothetical manifestations of what might be (Messick, 1979).

The task of item writing is greatly facilitated by the processes of implicit personality theory. These permit stable, reliable inferences between the trait definition and individual personality items. The evidence of the reliability of these inferences is compelling. In addition, the evidence indicates that item pools so constructed, even without empirical selection, possess certain levels of desirable empirical properties such as internal consistency and validity. Additionally, substantively relevant items survive cross validation far better than do those items only showing initial empirical validity (Goldberg & Slovic, 1967). I was so convinced of the ability of the average undergraduate student to generate personality items capable of yielding empirical validity that in 1971 I issued a challenge (Jackson, 1971) to pit the most elaborate empirical item selection strategy against the efforts of untrained individual item writers, each spending a limited time constructing a scale reflecting a particular definition. Results from Ashton and Goldberg (1973) and from my own study (1975) support the superiority of undergraduates over psychotechnicians. Figure 1 provides a comparison of the average true-score criterion variance accounted for (as reflected in a roommate rating) of individual student-constructed scales without item selection and similarly short empirically-derived scales from the California Psychological Inventory.

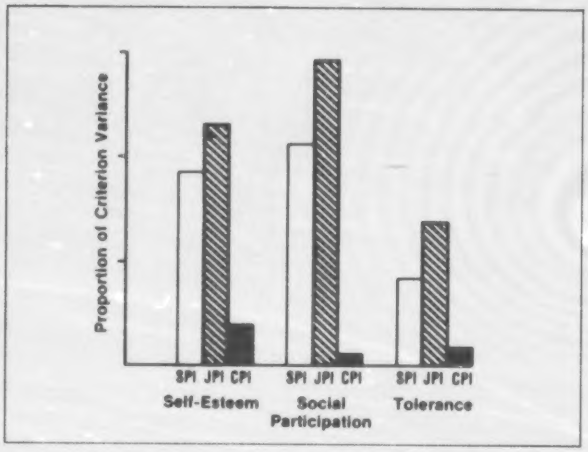


Figure 1. Proportion of criterion peer rating variance accounted for by three personality scale construction strategies: Student Personality Inventory (SPI); Jackson Personality Inventory (JPI); and California Psychological Inventory (CPI).

While some (Fiske, 1978) lament these results as reflecting poorly on the state of personality assessment, I prefer to view them as highlighting the powerful role of construct definitions in scale construction, so powerful, that with them the average undergraduate psychology student can turn in a very creditable performance.

But the conceptual convergence of personality trait definitions and item exemplars is not sufficient. If one is to hope for convergent and discriminant validity of personality scales later, it is necessary to foster these properties at the earliest stages of scale construction (Jackson, 1970). This should initially be undertaken judgmentally in scale definitions, item writing, the editorial process, and later by evaluating evidence from multivariate rather than from the traditional univariate item analysis. In fact, the recognition that each item reflects a great number of traits, albeit with unequal quality and precision, as well as representing a trait-method composite (Campbell & Fiske, 1959), suggests that the traditional total item score is likely to represent too imprecise and conglomerate a measure upon which to evaluate individual items. For this reason my recent test construction efforts have involved factor or components analyses in which item responses are evaluated in terms of their association with a particular orthogonal component score from which known linear sources of response bias and conceptually irrelevant traits have been removed statistically. These procedures permit a degree of statistical control over personality measures that is unattainable by any known experimental procedure. Indeed, the availability of large-scale computers permits additional refinements, such as the suppression of desirability response bias (Jackson, 1974; Neill & Jackson, 1970; Wiggins, 1973) and the development of algorithms for the minimization of redundancy in a set of measures (Neill & Jackson, 1976) by employing numerical procedures to evaluate the effects of different combinations of weights upon indices of scale redundancy. Procedures aimed at maximizing the content specificity of scales in a set of measures provide enriched opportunities for the set to demonstrate higher levels of prediction of a criterion than with more highly correlated sets.

To provide an illustration of the steps in the construction of a modern multi-scale device, I refer to the flow diagram contained in Figure 2 describing the steps involved in selecting items for a vocational interest survey (Jackson, 1976). Some of the steps summarized by a box in the diagram involve a number of sub-operations, which in turn could have been represented by additional flow diagrams.

The flow diagram is divisible into five general stages: (a) the identification and definition of relevant dimensions, item writing and editorial review; (b) administration, decomposition of data matrix into variance attributable to response bias and that due to content; (c) item analysis in terms of orthogonal component scores representing each putative dimension free from linear effects of response bias; (d) application of the algorithm for minimum redundancy item analysis; and (e) assembling and editorial review of final scales. In each of these stages a central concern was the identification of items and scales tapping independent dimensions of vocational interest related to the relevant construct both substantively and empirically.

My purpose in providing some detail about a construct-oriented approach to scale construction is twofold: first, it serves to illustrate the impact of this thinking on the development of measures, and second, it suggests that advances in this area may render our preconceptions about the substantive and psychometric properties of personality measures obsolete. Conclusions regarding the empirical validities of personality questionnaires developed 30 or 40 or even 50 years ago (Ellis, 1946; Guion & Gottier, 1956) simply may not apply to measures based upon adequately-defined constructs and abetted in their development by computer-based sequential strategies of scale construction. It is likely that through a series of successive approximations personality measures in the foreseeable future will evolve to the point where their links to constructs are solid, and their implications for reflecting behavior in the real world are known with greater certainty.

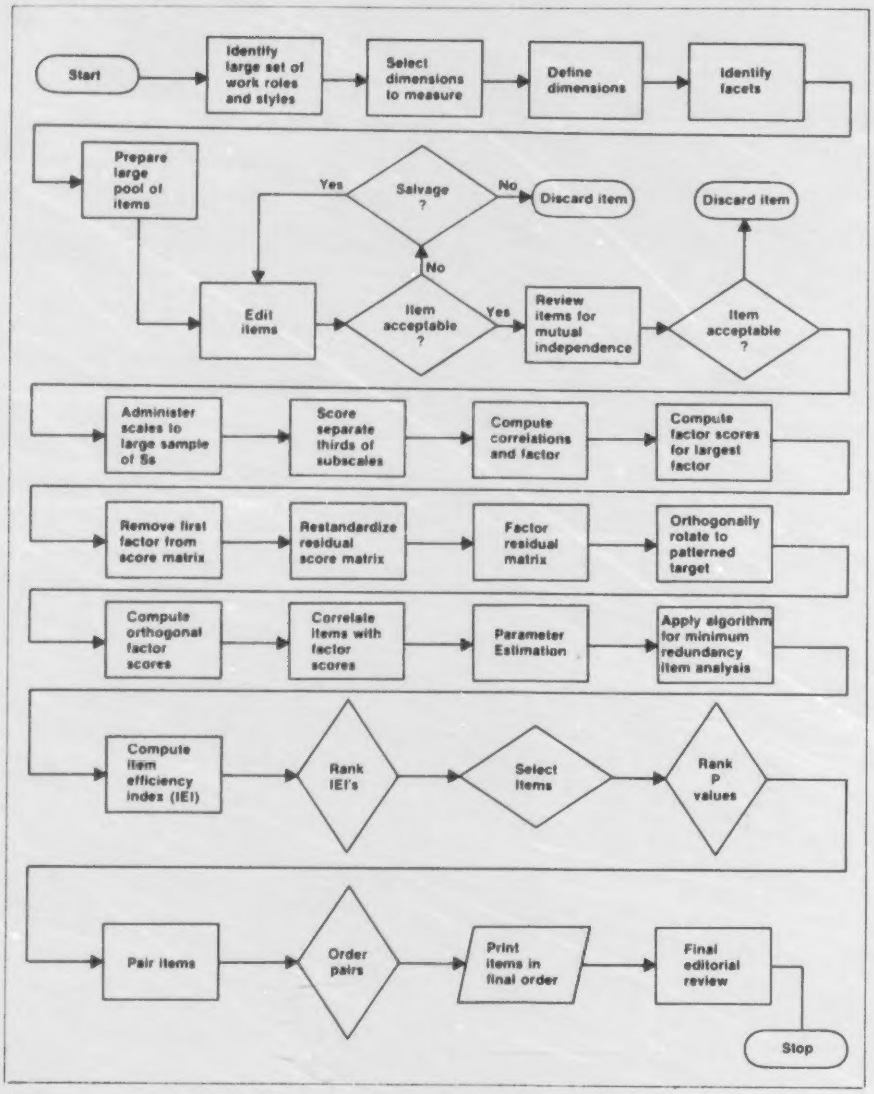

Start
Identify large set of work roles and styles
Select dimensions to measure
Define dimensions
Identify facets
Prepare large pool of items
Yes
Salvage ?
No
Discard item
Discard item
No
Edit items
Item acceptable ?
Yes
Review items for mutual independence
Item acceptable ?
Administer scales to large sample of Ss
Score separate thirds of subscales
Compute correlations and factor
Compute factor scores for largest factor
Remove first factor from score matrix
Restandardize residual score matrix
Factor residual matrix
Orthogonally rotate to patterned target
Compute orthogonal factor scores
Correlate items with factor scores
Parameter Estimation
Apply algorithm for minimum redundancy item analysis
Compute item efficiency index (IEI)
Rank IEI's
Select items
Rank P values
Pair items
Order pairs
Print items in final order
Final editorial review
Stop

Figure 2. Flow Chart for the JVIS Scale Development

Construct Validity and the Prediction of Behavior

Turning now from scale construction to the prediction of behavior or performance, I would like to focus upon issues relating to how a construct approach to personality might bear on the better understanding of criteria. I see five issues of major concern. These are (a) the issue of fakability; (b) reliability and the prediction of behavior; (c) the implications of trait vs. nomological validity and generalizability, (d) evaluation of reliability and validity of implicit personality theory; and (e) studies of personality factors in the employment interview.

Fakability

The actual or presumed fakability of personality measures is considered by many as sufficient reason for not considering this domain further in gathering information for selection or placement decisions. But this issue, like others, bears rational and empirical analysis. The susceptibility of some personality measures to positive faking is generally of one of at least three types: (a) the defensive or lie-scale variety, where a person presents an atypically and implausibly favorable representation; (b) general desirability, where the individual endorses content reflecting higher levels of consensually-regarded desirability; and (c) self-presentation in terms of a differentiated conception of a particular criterion group. With regard to the first, the lie scale variety, it is noteworthy that this response style is manifested with any frequency only when two conditions are simultaneously present: (i) instructions to provide an overall favorable impression; and (ii) a defensiveness or lie scale to elicit this behavior. In the absence of instructions to fake and a scale to measure it, this bias occurs relatively infrequently, and is best interpreted as a personality style (Crowne & Marlowe, 1964) rather than as deliberate faking per se. The incidence of this and other types of faking, even in highly charged selection situations, is considerably lower than one might expect. Norman (1973) has outlined techniques for detecting and controlling for this type of faking if and when it occurs in selection decisions.

Individual differences in responding desirably are undeniably present in personality questionnaires, and the mean level of desirability responding is sometimes, but not always, affected by faking instructions. Both the extent of the response style and of a scale's sensitivity to situations evoking it may be drastically reduced by choosing items in scale construction in a narrow neutral, balanced range of desirability. Even when it occurs in situations where subjects are instructed to fake, or respond desirably, the rank ordering of individuals across straight-take and faking or desirability conditions tends to remain stable (Kusyszyn & Jackson, 1968; Hoffman, 1968). Furthermore, contrary to widely held belief, Holden and Jackson (1979) have found that subtle items are less valid than obvious items, both under "straight take" and faking instruction. Differences in desirability responding, whether due to response style or to motivated distortion, may, of course, be statistically removed from other scores, but the possibility that this response style might be related to a criterion measure should not be overlooked.

Self presentation in terms of a particular criterion group involves a kind of role playing which, although quite possible, occurs less frequently than might be expected. Indeed, in two studies undertaken with medical school applicants who were under intense pressure to gain admission, I (Jackson, 1964) found in the first that students under "straight take" instructions and under a variety of dissimulation instructions, including those to give responses to provide the most favorable impression for admission to medical school, failed to elicit significant differences for any of the 20 Personality Research Form scales between instructional groups. In a second study, a vocational interest survey given to the same students before and sometime after admission to medical school showed substantial individual stability in the rank ordering of scale scores (Jackson, 1977). If anything, mean scale scores shifted from somewhat lower to higher congruence with the physician profile after admission. My hunch is that respondents regard the response process as an integral manifestation of the self (Kuiper & Rogers, 1979), one that is not readily amenable to change. When the self is not implicated, and they are asked merely to judge the characteristics of a person in a particular criterion group, this is accomplished with alacrity (Rothstein & Jackson, 1979).

The degree to which role-playing occurs, and the methods, if any, by which it should be controlled, will remain a continuing area for empirical analysis. It is not, in my view, as serious or as formidable a problem in test fairness as is the issue of test-wiseness in aptitude and achievement testing.

Reliability and the Prediction of Behavior

Turning now to the question of reliability, it is frequently asserted that it is not reasonable to expect that personality scales will demonstrate internal consistency reliability to the same level as those obtained for general aptitude measures. Indeed, much of the published evidence might appear to support this assertion. But there are two important reasons for treating such observations with skepticism. First, it is well to recognize that many personality scales have been constructed using ad hoc external criteria in such a way that evidence of reliability is neither fostered nor expected (Gough, 1957). The situation with respect to construct-oriented scales bears examination. Second, personality scales are typically short—the assessment specialist is not afforded the privilege of suspending the well-known relationship between reliability and test length just because the focus has shifted from ability to personality. Where construct-oriented approaches to scale construction are used, evidence for reliability is found. In the study of novice item writers (Jackson, 1975), each student produced one personality scale with 16 items, unselected by virtue of any empirical data. The median coefficient alpha was .51, but before complaints are voiced that these values are only moderate, remember that they could easily have been combined into 100-item scales with substantially higher reliabilities. In a more elaborate scale-construction effort, I (Jackson, 1974) obtained a median coefficient alpha of .92 for 20 personality scales comprised of unselected items numbering about 100 for each scale. Even the often-quoted hoary Hartshorne and May (1928) study on the supposed specificity of honesty yields to an analysis based on the even earlier developments of Spearman and Brown. When their seven by seven intercorrelation matrix of honesty behavior items is projected to a more typical length of 50 items, the resulting reliability is estimated to be .97 (Jackson & Paunonen, 1980, cf. Epstein, 1979). Honesty measures would thus appear to be at least on a par with, say, scholastic and aptitude measures with respect to reliability.

Nor can the relationships between reliability and validity from classical test theory be safely ignored in the personality area. Moskowitz and Schwarz (1979) have recently reported that validity increased linearly and substantially as a function of the number of behavior samples observed and the number of judges. Moskowitz (1979) has also reported evidence that broad constructs based on multiple and varied behaviors and across diverse situations will better predict a referent than will data sampled in only one situation. The results from the validity of scales developed by naive item writers also support the view that more reliable scales yield higher validities.

Authors of general-purpose personality questionnaires have been sensitive to the constraints imposed by the bandwidth-fidelity dilemma (Cronbach and Gleser, 1965) and have tended to proffer many relatively short scales, all of which could be administered in an hour or less. The result often is broad bandwidth, usually at the expense of fidelity. Where the utility of high reliability and validity of single scales is substantial, assessment specialists should consider lengthening total testing time and/or selecting a smaller number of personality scales conceptually focussed with regard to criteria.

Trait and Nomological Validity

A useful distinction may be made in construct validation between trait validation and nomological validation (Campbell, 1960; Cronbach & Meehl, 1955; Messick, 1979). Putative measures of traits are properly evaluated in terms of their convergence with distinct, heteromethod, heterosituational measures of the same trait, and in terms of their relative independence from measures of conceptually distinct traits. Nomological validity refers to the network of theoretical expectations regarding the links in different combinations between abstract traits, concrete exemplars of abstract traits, and complex criteria and behaviors.

The nomological aspect of construct validity has been criticized by behavioristically-oriented psychologists first, because it is felt that statements about nomological networks are premature with the present state of formal theorizing (Campbell, 1960) and second, because, philosophically, such inferred entities are viewed with suspicion. Abstract constructs with clear referents do not make me uncomfortable, but I freely admit that if we only had formal theory upon which to base our personality scale construction and interpretation efforts, they would hardly get off the ground. Fortunately, there is a network of trait and behavioral relationships known to researchers in person perception as lay or implicit personality theory. If it could be established that judges given limited information about a person could reliably and accurately predict that person's probability of behaving in a certain way in different situations, such findings would have an important bearing on the use of implicit personality theory in explicating the nomological network, both to cast light upon formal properties of this network, and to explicate the personality construct properties of criterion behaviors. The explication in construct terms of criterion measures is at least as important to validation as is the construct validation of predictor measures (Messick, 1979; Tenopyr, 1979). There is some evidence that such judgments do show stability, high inter-judge reliability, and validity in predicting behavioral relationships in laboratory studies. I would like to share with you results from a few studies offering promising support for the utility of human judgment.

Evaluations of the Reliability and Validity of Implicit Trait Relationships

Judgments and predictions of behavior are necessarily probabilistic. Persons labeled as aggressive by virtue of a high personality scale score are not always aggressive. Given appropriate instigation, the probability of aggressive behavior should be higher. For this reason the majority of our studies seeking to explore properties of implicit personality theory have relied on judgments of probability, frequency, or certainty. In what was the first of a number of multidimensional scaling studies revealing very high reliabilities for group consensus judgments, Walters and Jackson (1966) found that the projections on the major dimensions for two randomly-assigned groups judging the joint probability in people of trait names correlated in the high nineties, rivaling results from color studies. In a further study employing more concrete personality items together with characterizations of persons high or low on the personality traits of Dominance, Autonomy, and Impulsivity two randomly assigned groups of judges whose multidimensional scaling solutions were independently analytically rotated yielded projections on three dimensions which correlated .99, .99, and .98 respectively (Jackson, 1971). A variety of other findings indicate that there is a high degree of consensus for judgments of the joint probability of occurrence of exemplars of personality traits, even when these are presented in non-verbal, imaginal terms. Paunonen and Jackson (1980) in an experimental study of personality impression formation employed verbal and non-verbal materials in a balanced design both to provide information about target persons and to secure ratings. Each target individual was selected to be similar to one of two psychological types isolated and replicated by multivariate clustering procedures of large samples of personality profiles generated from the Personality Research Form. Information was drawn from four scales but predictions of probability of occurrence of the behavior was related to 17 scales. An example of the simple non-verbal materials used is contained in Figure 3. Results indicated a high degree of consensus and accuracy in predicting behavior relevant to each of a number of personality dimensions represented by the profile presented in Figure 4 about target behavior regardless of whether information was provided judges imaginally or verbally or whether judgments were made in either mode. Furthermore, the profile resulting

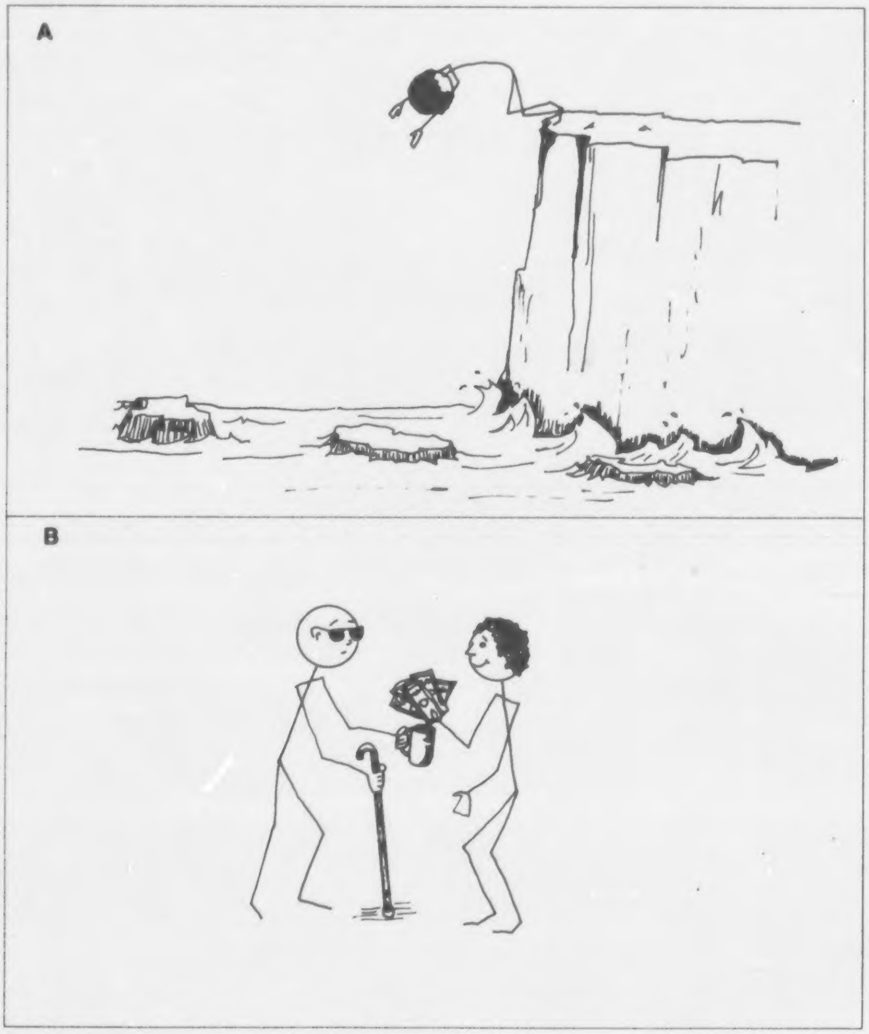


Figure 3.

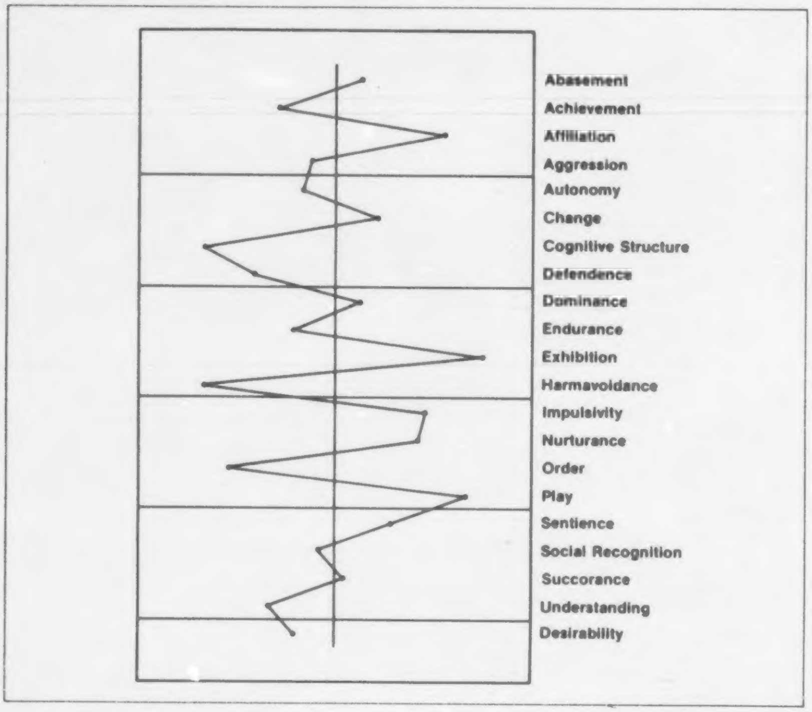


Figure 4.

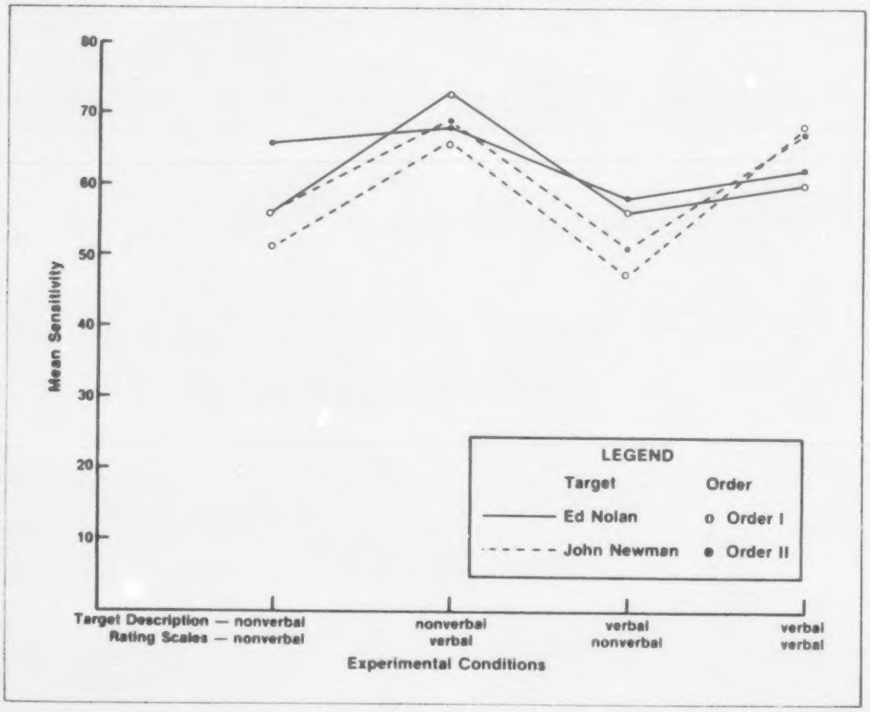


Figure 5.

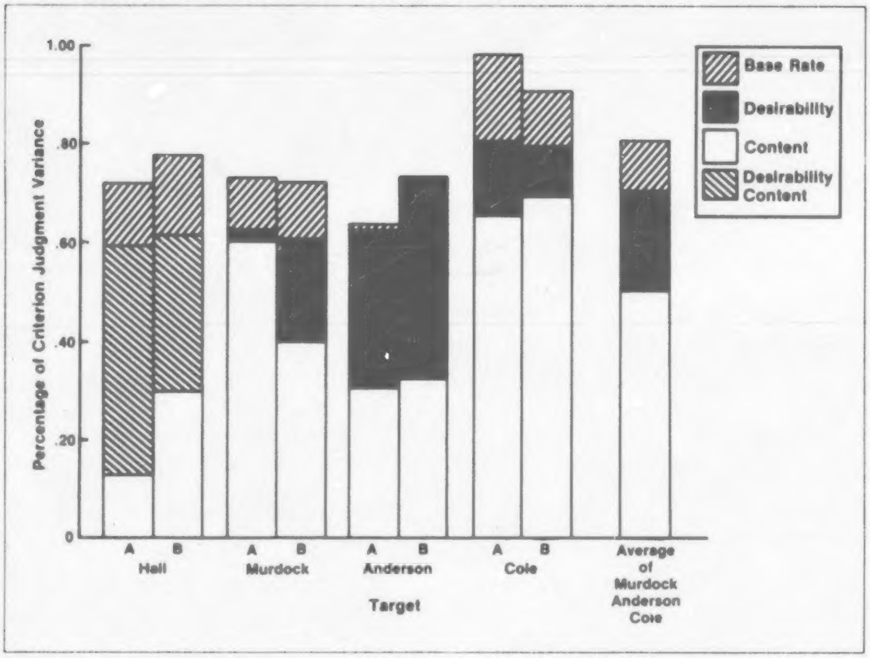


Figure 6. Percentage of inferential judgment variance attributable uniquely to each of the three components, content, base rate, and desirability.

from probability judgments corresponded closely to the personality profiles of the target persons' clusters (Figure 5). Reed and Jackson (1977) in an extensive series of studies selected seven persons exemplifying each of four types of persons again identified by multivariate clustering. Judges, given limited information in the form of brief personality sketches, yielded multiple correlations with actual pooled item responses of target persons over a broad range of content ranging from .83 to .95 when judged personality content, judged desirability, and judged base rates were taken into account (Figure 6). My interpretation of these studies is that judges, given limited but relevant information about target persons, are capable of predicting with relatively high accuracy the concrete response behavior of a (reliable) set of targets, even for behaviors which are related only minimally with the information at hand.

Studies of Personality Factors in the Employment Interview

It would appear that there is one almost universally applied personality assessment device in use in employee selection, one often not recognized as such, namely, the employment interview. Indeed, leading "how-to" texts about selection interviewing (e.g., Fear, 1978) speculate on required personality characteristics for different jobs and how these might be judged in a candidate.

We have conducted four related studies (Jackson, Peacock & Smith, 1980; Rothstein & Jackson, 1979). I would like to share with you some of the results obtained in one of these (Rothstein & Jackson, 1979). Based on factor analyses of Personality Research Form and SVIB occupational scales (Siess and Jackson, 1970), predictions were made about the personality characteristics of engineers and accountants. Empirical results suggested that engineers were high in PRF scales for Endurance, Achievement, and Understanding, and low in Succorance and need for Social Recognition. Incidentally, this is also the empirically-obtained profile for some 200 data processing personnel. Accountants were high in Order and Cognitive Structure and low in Impulsivity, Change, and Autonomy. Based on this information, we wrote scripts for interviews in which candidates presented themselves as possessing the personality characteristics of either the engineer or of the accountant. Embedded in the interview were self-referent statements drawn from relevant scales from the PRF. An experienced interviewer and college student volunteer helped us prepare tape recorded examples of employment interviews. Figure 7 shows results when judges reported their ratings for persons with each type of personality for each job. It will be noted that ratings for the personalities relevant to the job were very much higher than were those whose personalities were relevant to the alternative job. Furthermore, when judges were asked to predict the personality item responses of each of the candidates, the overall group consensus reliabilities for both the accountant and engineer applicants was .99. (It was slightly lower when only applicants whose personalities were incompatible with the job were considered.) These results parallel those of Blair and Jackson (1978), who found similar levels of reliability for group consensuses in judging the vocational interests across the 34 scales of a Vocational Interest Survey of six occupational groups. But not all judges show the same degree of sensitivity. Figure 8 depicts the distribution of judges in approximating the group consensus judgment regarding applicants. It will be observed that while most judges were sensitive to the patterning of target behaviors, a few operated at chance levels. These results imply that, as with other abilities, the

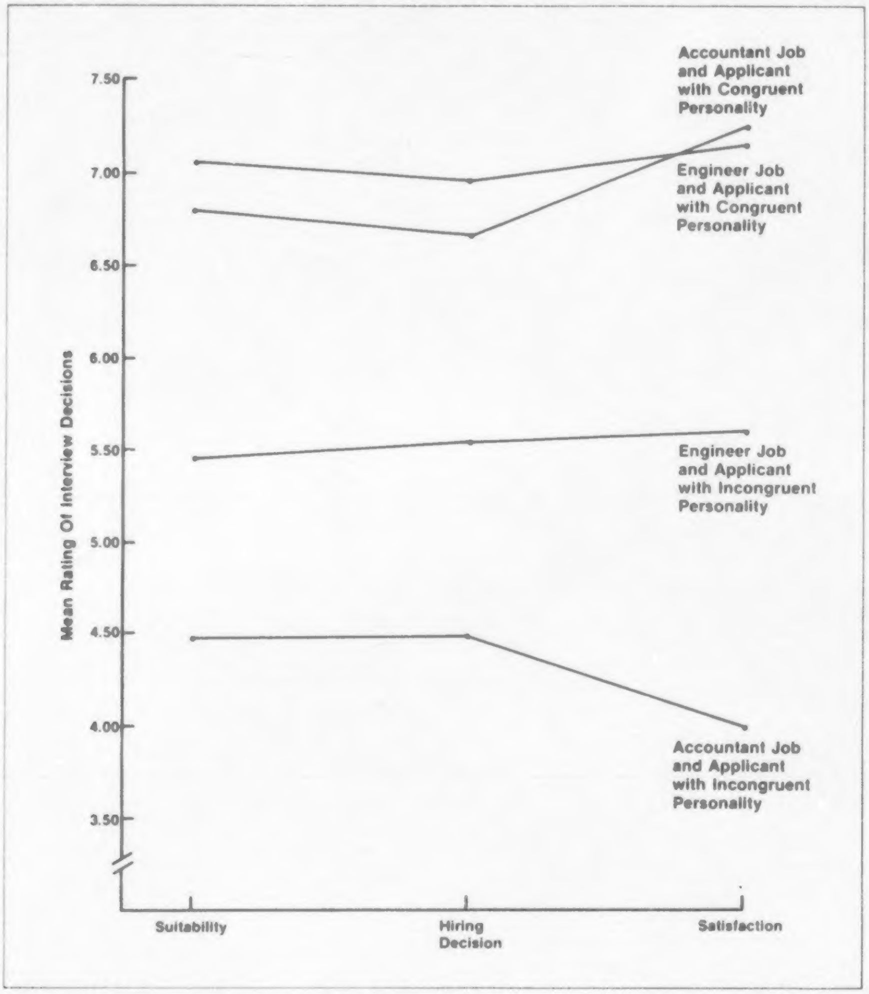


Figure 7. Mean selection decisions in the Applicant Target × Job Category interaction.

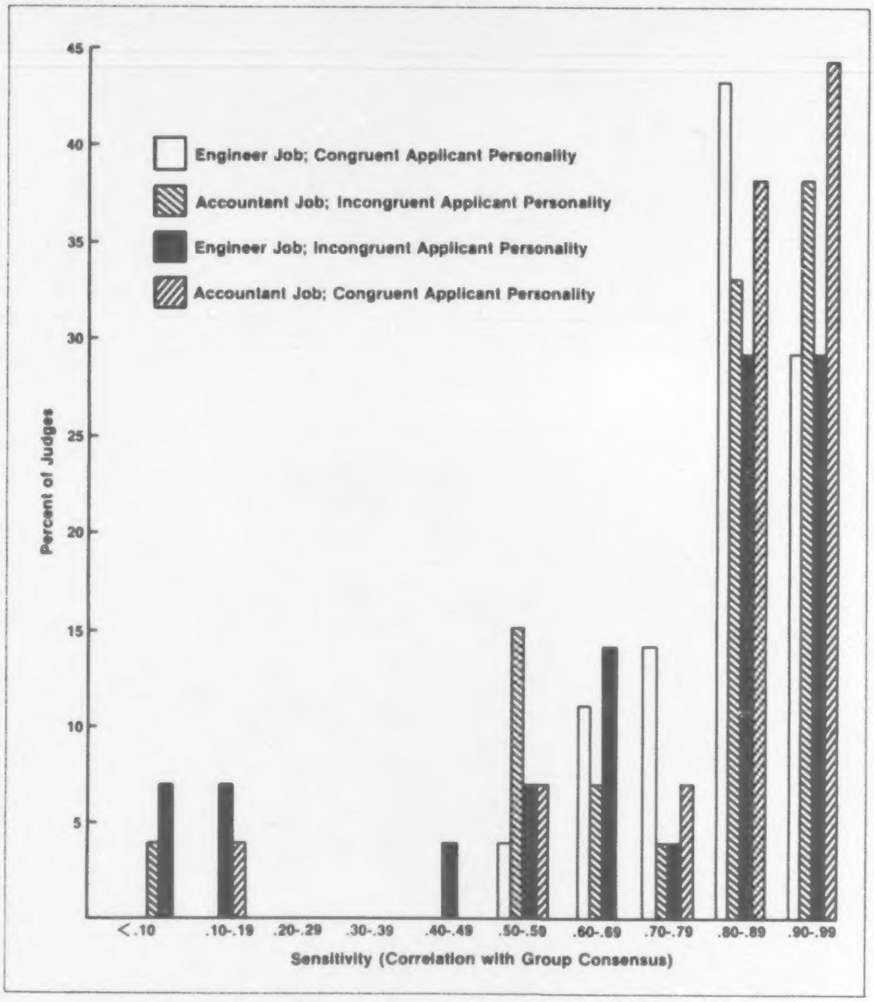


Figure 8. Frequency distributions of sensitivity, as estimated by the correlation between item group consensus scale values for each condition and an individual's judgments of target likelihood of endorsement.

ability to judge personality sensitively is not uniformly distributed among judges. Training might help (cf. Strasburger & Jackson, 1977).

Overview

I hope that the above results have served to illustrate some of the potential for a construct-oriented approach to personality scale construction and assessment, and for the use of implicit personality theory as developed in the collective wisdom of judges for explicating the implicative structure of behavior, of situational or job requirements, and of the relation between the two. Obviously much more work is needed to evaluate and compare conceptual analyses of prediction situations with alternatives.

References

Ashton, S. G. & Goldberg, L. R. In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by the experts, novices, and laymen. *Journal of Research in Personality*, 1973, 7, 1-20.

Blair, N. L., & Jackson, D. N. Inferring vocational interest. Research Bulletin 435, Department of Psychology, The University of Western Ontario, London, Canada, 1978.

Braithwaite, R. B. *Scientific explanation.* Cambridge: Cambridge University Press, 1953.

Campbell, D. T. Recommendations for APA test standards, trait, or discriminant validity. *American Psychologist,* 1960, *15,* 546-553.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81-105.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology,* Chicago: Rand McNally, 1976.

Cronbach, L. J., Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, *52,* 281-302.

Crowne, D. P., & Marlowe, D. *The approval motive: Studies in evaluative dependence.* New York: Wiley, 1964.

Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. *Annual Review of Psychology,* 1979, *30,* 477-525.

Edwards, R. C. Personal traits and "success" in schooling and work. *Educational and Psychological Measurement,* 1977, *37,* 125-138.

Ellis, A. The validity of personality questionnaires. *Psychological Bulletin,* 1946, *43,* 385-440.

Epstein, S. The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology,* 1979, *37,* 1097-1126.

Fear, R. A. *The evaluation interview: Predicting job performance in business and industry* (2nd ed.). New York: McGraw-Hill, 1978.

Fiske, D. W. *Strategies for personality research.* San Francisco: Jossey-Bass, 1978.

Goldberg, L. R. & Slovic, P. The importance of test item content: An analysis of a corollary of the deviation hypothesis. *Journal of Counseling Psychology,* 1967, *14,* 462-472.

Gough, H. G. Manual for the California Psychological Inventory. Palo Alto: Consulting Psychologists Press, 1957.

Guion, R. M., & Gottier, R. F. Validity of personality measures in personnel selection. *Personnel Psychology,* 1956, *18,* 135-164.

Harthshorne, H., & May, M. A. *Studies in the nature of character: Vol. 1. Studies in deceit.* New York: Macmillan, 1928.

Hempel, C. G. *Fundamentals of concept formation in empirical science.* Chicago: University of Chicago Press, 1952.

Hempel, C. G. *Aspects of scientific explanation.* New York: Free Press, 1965.

Hoffmann, H. Performance on the Personality Research Form under desirable and undesirable instructions: Personality disorders. *Psychological Reports,* 1968, *23,* 507-510.

Holden, R. R. and Jackson D. N. Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 459-468.

Howell, M. A. Personal effectiveness of physicians in a federal health organization. *Journal of Applied Psychology,* 1966, *50,* 451-459.

Jackson, D. N. A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology,* Vol. 2. New York: Academic Press, 1970.

Jackson, D. N. The dynamics of structured personality tests: 1971. *Psychological Review,* 1971, *78,* 229-248.

Jackson, D. N. *Personality Research Form Manual.* Rev. Ed. Port Huron, MI: Research Psychologists Press, (2nd ed.), 1974.

Jackson, D. N. The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational and Psychological Measurement,* 1975, *35,* 361-370.

Jackson, D.N. *Jackson Personality Inventory Manual.* Port Huron, MI: Research Psychologists Press, Inc., 1977.

Jackson, D. N., & Paunonen, S. V. Personality structure and assessment. *Annual Review of Psychology,* 1980, *31,* 503-551.

Jackson, D. N., Peacock, A. C., & Smith, J. P. Impressions of personality in the employment interview. *Journal of Personality and Social Psychology,* 1980, in press.

Kuiper, N. A., & Rogers, T. B. Encoding of personal information: Self-other differences. *Journal of Personality and Social Psychology,* 1979, *37,* 499-514.

Kusyszyn, I., & Jackson, D. N. A multimethod factor analytical appraisal of judgment and response methods in personality assessment. *Educational and Psychological Measurement,* 1968, *28,* 1047-1061.

Landy, F. J. The validity of the interview in police officer selection. *Journal of Applied Psychology,* 1976, *61,* 193-198.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports,* 1957, *3,* 635-694.

Margenau, H. *The nature of physical reality.* New York: McGraw-Hill, 1950.

Messick, S. J. Test validity and the ethics of assessment. Paper presented at the meeting of the American Psychological Association, New York, September, 1979.

Moskowitz, D. Analyzing personality constructs for internal consistency and transsituational generality. Paper presented at the meeting of the American Psychological Association, New York, September, 1979.

Moskowitz, D., & Schwarz, J. C. In search of better personality measures: The validity of ratings and behavioral observations. Paper presented at the meeting of the Eastern Psychological Association, Philadelphia, April, 1979.

Neill, J. A., & Jackson, D. N. An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement,* 1970, *30,* 647-661.

Neill, J. A., & Jackson, D. N. Minimum redundancy item analysis. *Educational and Psychological Measurement,* 1976, *36,* 123-134.

Norman, W. T. Personality measurement, faking, and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology,* 1963, *47,* 225-241.

Paunonen, S. V., & Jackson, D. N. Nonverbal trait inference. *Journal of Personality and Social Psychology,* 1980, *37,* 1645-1659.

Reed, P. L., & Jackson, D. N. Personality measurement and inferential accuracy. Research Bulletin No. 419, Department of Psychology, The University of Western Ontario, London, Canada, 1977.

Rothstein, M., & Jackson, D. N. Personality factors in the employment selection interview. *Journal of Applied Psychology,* in press.

Siess, T. F. and Jackson, D. N. Vocational interests and personality: An empirical integration. *Journal of Counselling Psychology*, 1970, *17*, 27-35.

Strasburger, E. L., & Jackson, D. N. Improving accuracy in a clinical judgmental task. *Journal of Consulting and Clinical Psychology*, 1977, *45*, 303-309.

Tenopyr, M. L. *Discussion*. Revising the 1975 APA Test Standards, presented at the American Psychological Association, New York, September, 1979.

Walters, H. A., & Jackson, D. N. Group and individual regularities in trait inference: A multidimensional scaling analysis. *Multivariate Behavioral Research*, 1966, *1*, 145-163.

Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.

DISCUSSION

Hunter: I would like to corroborate one thing you said. For many years, because of my early training I felt some guilt in using thought in the construction of scales and relying on judgment in reading items and figuring out what they should measure, so I kept track very carefully of the difference between statistical fit and fit in terms of content for items. I found exactly what you did. Consistently on cross-validation, content was almost always a better indicator of an item's validity than were the statistical properties from, say, factor analyses.

Jackson: I admit I am glad that's been your experience. I asserted at one point that all the so-called subtle items on the MMPI were simply there due to chance, but nobody has ever cross-validated the MMPI, to my knowledge, so we don't have the answer to that.

Thorndike: I can report on researchers from Queens College who did, and obtained shrinkage which was fairly directly proportional to the subtlety of the item.

Jackson: Thank you. I was not aware of such a study, but I am glad to be aware of one now.

Norris: I notice on Figure 2 you don't have a decision box to judge whether or not items are acceptable after the test has been given for the first time. Was that intentional? I would think you could use information from people's response to decide whether or not items were acceptable. That would be my assumption.

Jackson: I apologize if I don't have a decision box. Essentially the items are ranked, using an overall statistical index. We did the item analysis twice, optimizing different properties each time and coming to a smaller and smaller number of items. But there should be a decision box there. The decision was to draw the line at a certain point, rather than some place else, when the items are ranked in terms of some statistical property.

Norris: You discard items, then?

Jackson: Yes, items are discarded at all levels. And this is actually a highly abbreviated description of what actually was done.

Gulliksen: You mentioned correlation. Was it an inter-item correlation?

Jackson: The items are correlated in this particular instance with a factor score derived from combining three subsets of items from the same scale.

Gulliksen: You say compute correlations and factor?

Jackson: I'm sorry. That's another part. That was not item correlation. In this case it was correlations among partials of items.

Gulliksen: Then you would have only three correlations, in fact?

Jackson: No, each one of a number of scales, 20 or more scales. So what I actually dealt with here were four sets of four hundred by twenty-two matrices, with four hundred subjects and twenty-two scales, and something over 2,000 items in this analysis.

Thorndike: When you remove the first factor, you are assuming the first factor is essentially garbage, socially desirable, or something similar?

Jackson: These were vocational interest items I happened to be talking about here. I didn't exactly expect it quite as strikingly as I found it. I found a third of the variance was attributable to the tendency to like everything. It didn't matter what, you know—carrying garbage, cutting diamonds, anything. And that might seem just like a rather trivial psychometric problem to some, at least, but this problem has important implications for counseling. I believe in returning results to respondents. I explained that it was highly tentative and all of that. But some people came to me with profiles where everything was at the 99 percentile in this initial screening, and other people showed everything at below the 15 percentile. And that's the sort of thing we get with response bias. I did not want to include that in the total score, against which I selected items.

Question: This would be the traditional acquiescence?

Jackson: I am reluctant to use that word. It usually starts an argument. It was a tendency to like a lot of things.

Bond: Do you mean you didn't find the typical bias against outside work?

Jackson: Some people were selective in terms of general and specific item content, some people liked everything, some people didn't like everything. And that's the reason in this case that I used the forced choice form. There was really no alternative.

Horn: What is the average degree of independence of the scales?

Jackson: With regard to the reliabilities? From the published personality research form, the average absolute scale correlation is in the vicinity of .2. And the reliability—there are different ways to compute it, but in the original long form the median reliability was .92. In the short final form, the reliabilities were in the vicinity of .8 on each of the 20 scales. Actually, I can generate 10 factors from these vocational interest scales, replicated across genders, which have a median reliability of .92, so that there is, I think, some degree of scale independence.

Horn: You defend that on grounds of psychological utility?

Jackson: Well, there are a number of reasons for preferring independent scales. For one, from a construct point of view, it means that the scales do not share convergent variance in a sense. From the point of view of convergent-discriminant validity, we are better off having independent predictors. In the limiting case where separate tests measure only one factor, discriminant validity is not possible.

Horn: I thought you were saying in effect there are not 20 or 22 dimensions here, as a matter of fact, if you really start looking at the degree of independence in terms of the overall set and, therefore, you would want to keep 22 dimensions because they had psychological meaning.

Jackson: Every item showed substantially a higher correlation with the component on which it was supposed to be represented than it did with any other component. I have cross-replicated that for the Personality Research Form where strong evidence for separate dimensions was obtained. We even did it in the French and German translations. And in both cases it was the same; there was a very strong evidence that each scale defined a separate dimension. Item factor analyses also support this interpretation.

Sternberg: It seems what you are doing might have generality beyond the study of personality. One of the points Carroll made was that one thing we want to do in the intelligence field is to relate our theoretical notions to more everyday terms people use in understanding the term "intelligence." It seems to me a lot of methodologies you are proposing here could be used to understand people's interpretation of intellectual performance, and in that way eventually perhaps form some kind of basis for the construct validity of intelligence. In other words, just substitute "implicit personality" theory for a theory of intellective behavior using these same methods.

Jackson: I thought about this, actually. I don't know if it was in anticipation of what you were going to say or what. But my impression is that this kind of implicit theory would be much more difficult. It might actually take experts.

My position is that with personality, most people have substantial experience observing other people's behavior every day. Their intellectual processes are often more private and more difficult to discern. But on the other hand, for subject matter experts, I think the kind of more formal explication of what we are talking about via a component or process or factor analysis probably merits more explicit empirical treatment of judgments rather than more informal discussion with subject matter experts.

Sternberg: Alternatively, what you might do is to carry it one step further and do both. We are doing that in a slightly different context. And that is, compare the expert's judgments with people-in-the-street's judgments, and see what kinds of correspondence you get. You could do a similar thing for personality and you might find, as you suggested, that there would be more difference in the intellectual domain than the personality domain.

Jackson: That would be interesting.

Gulliksen: The peer rating in Figure 1, does that mean students rated the other students? Who were the peers?

Jackson: This was a roommate who had lived with the target person for five months or more.

Gulliksen: These were a group of students?

Jackson: They were students, volunteers. Of course, the reliability of rating is a function of the number of judges, and we just had one here.

SAMUEL MESSICK

CONSTRUCTS AND THEIR VICISSITUDES IN EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*

Each of the classical views of test validity carries with it a tacit prescription for test development. *Criterion-related validity* implies an empirical strategy whereby items are selected that significantly discriminate between relevant criterion groups or that maximally predict relevant criterion behaviors. Such tests are usually evaluated and sometimes revised in terms of the degree to which the items retain their predictive properties under cross-validation, the degree to which item responses display homogeneity or internal consistency, and occasionally the degree to which item content is coherent enough for sensible interpretation.

Content validity implies a rational strategy whereby a particular behavioral domain of interest is identified, usually by reference to curriculum objectives or task requirements or job characteristics. The boundaries and various facets of the domain are specified, and items are constructed to cover those facets in some representative fashion. Such tests are usually evaluated and sometimes revised in terms of the degree to which item responses display homogeneity and the degree to which scores relate to relevant external criteria or discriminate between relevant criterion groups.

Construct validity implies a joint convergent and discriminant strategy entailing both substantive coverage and response consistency in tandem. The boundaries and facets of a behavioral domain are specified, but in this case as delineated by a theory of the construct in question. Items are written to cover that domain in some representative fashion, as with content validity, but in this approach the initial item pool is deliberately expanded to include items relevant to competing theories of the construct, if possible, as well as items theoretically irrelevant to the construct (Loevinger, 1957). Item responses are then obtained, and items are selected that exhibit response homogeneity consistent with the focal theory of the construct but that are distinct from theoretically irrelevant items or exemplars of competing theories. Both convergent and discriminant evidence are thereby obtained at the stage of test construction, and a miniature test of the construct theory is afforded at the item level. An efficient elaboration of this strategy is to develop measures of two or more distinct constructs at the same time and to administer the combined item pool along with measures of anticipated method contaminants such as social desirability. Provisional scale scores are then derived for each construct, and items are retained that correlate more highly with their purported construct than with other construct or response style scores. Thus, item selection is based on convergent and discriminant evidence conjointly, while at the same time method contaminants are suppressed (Jackson, 1971).

Tests developed by the construct strategy are usually evaluated and sometimes revised by appraising the extent to which the items included in and excluded from the test can be accounted for in terms of the construct's theory, which Loevinger (1957) called "substantive validity," and the extent to which the test relates more highly to different methods for measuring the same construct than it does to measures of different constructs assessed by the same method, which Campbell (1960) called "trait validity." The test's correlations with external measures of other constructs and with particular criterion behaviors bear on "nomological validity" (Cronbach & Meehl, 1955), which refers in the main to the validity of the construct as an integrative hypothesis. In the construct approach the test is not to be revised, of course, to heighten its relation to any one of these external measures, no matter how criterial it might appear in a particular applied setting, lest the test thereby become biased as a measure of the construct and consequently distort empirical estimates of the lawful connections with other constructs in the theoretical or nomological network.

Just as different views of test validity imply different strategies of test development, so do different views of mediating variables imply different strategies of measurement and interpretation. We will next consider various points of view about the import of test behaviors in relation to non-test behaviors, especially those interpretations of relationship ranging from the use of operationally-defined intervening variables, to the use of constructs as inductive summaries having only "systemic existence" in the theoretical system of the investigator, to the use of constructs as inferred entities having supposed real existence outside the mind of the psychologist (Beck, 1950). Note that this next section of the paper is not to be a discussion of constructs in the philosophy of science, but of constructs in the working worldviews of psychologists. A salient point of contention in contemporary psychology is not so much whether inferred entities are real, but which inferred entities are real, in particular whether dispositional traits are real attributes of persons or whether the behavioral consistencies they refer to are instead elicited and maintained by environmental factors, which in turn are the real entities. In intermediate views, some behavioral consistencies are referenced to traits, some to situational factors, and some to interactions between them, in various and arguable proportions (Bowers, 1973; Ekehammar, 1974). What is at issue in this discussion of the reality of traits is not metaphysics but methodology, for existential tenets tend to fix the locus of empirical inquiry and influence not only how we proceed but what we perceive.

---

*I wish to thank Lee Cronbach for his stimulating comments on the presentation, which led to some clarifying revisions and emendations. I also wish to thank Walter Emmerich for suggesting some important refinements in the development of the argument.

We will then briefly consider the role of assessment models, which are used either explicitly or implicitly in the measurement of a construct, in relation to the theory of the construct, examining the kinds of difficulties that arise when the two are not well coordinated. For present purposes, an assessment model is considered to include both the scoring or measurement model relating item responses to test score and the conditions of testing that influence score interpretation, such as degree of standardization, speededness, norm vs. criterion referencing, group vs. individual administration, and intended use. In a final section, we will view the assessment models as implicit psychological theories requiring either unified treatment within the nomological network of the construct measured or the development of collateral theories coordinating issues of form, structure, method, and context with the substantive interpretations of the construct.

Logical Positivism and Constructive Realism in Construct Theory

A variety of viewpoints about the import of test behaviors in relation to non-test behaviors will be addressed in this section in a necessarily telescoped fashion. First, we will examine the view that behavioral consistencies are manifestations of real traits in contrast to the position that consistency in behavior is instead attributable to real stimulus contingencies in the environment. Next, we will consider a view that summarizes behavioral consistencies in terms of a set of theoretical constructs and the lawful relationships among them, with meaning being grounded in the theoretical network rather than in the real world. Finally, we will examine the viewpoint that behavioral consistencies are indeed manifestations of real traits but are only understood in terms of constructs that summarize their empirical properties in relation to other constructs in a theoretical system. Figures used to facilitate the discussion in this section are so highly schematized that they may blur more issues than they illuminate, so they should be taken lightly for whatever light they shed.

Personal Causation or Environmental Control

Figure 1 presents a schematic portrayal of the psychological realist's viewpoint. Some characteristic of the person—an attribute or process or disposition, referred to here as a "trait" for short—*causes* both the test and the non-test behaviors, which are mutually related by virtue of common or related causes.

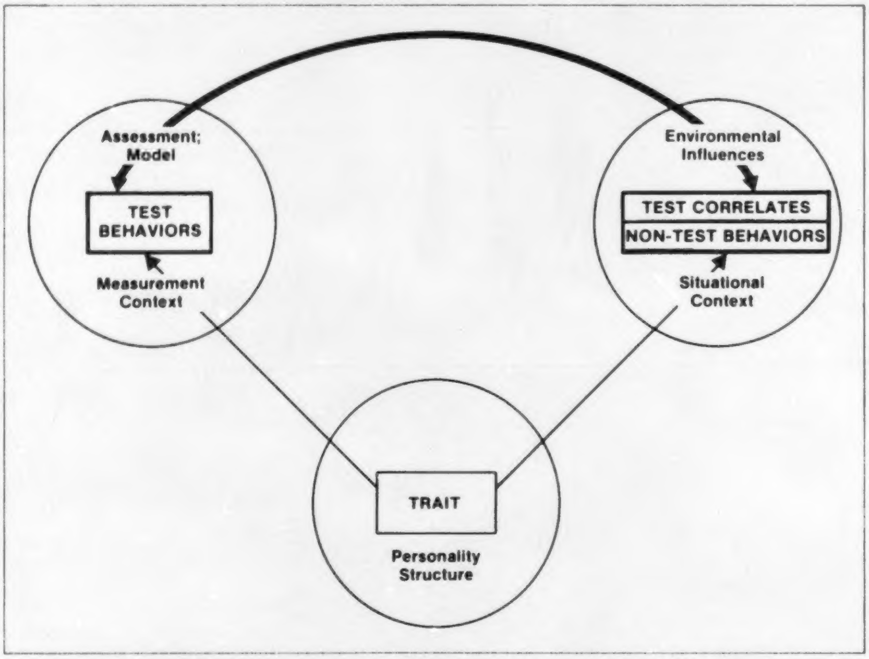


Figure 1

Some adherents of this position would likely stop short of attributing causality to traits, but rather would consider consistencies in test and non-test behaviors to be *manifestations* of traits, which organize behavior or otherwise produce response consistencies. For our purposes, the test behaviors may be either consistencies in response to items of a single test or, more usually, consistencies in response to different methods of measuring the same focal variable. At issue in relation to these test behaviors are both non-test behaviors and test scores for other nonfocal variables. The boxes enclosing the test behaviors and the non-test behaviors in the figures are darkened, as is the curved line of relationship between them, to emphasize that these are the primary *observables* under discussion. The non-test behaviors are surrounded by a situational context of environmental influences that can modulate empirical relationships, as can the measurement context surrounding the test behaviors. Furthermore, the focal trait is embedded in a personality structure implicating other related traits, which complicates the picture enormously. Thus, for example, one motive or need may operate in the service of another motive or need in a particular situation, or one characteristic of the person, such as defensiveness, may interfere with the measurement of other characteristics of the person in a given measurement context. The contemporary psychologist who comes closest to this realist position is probably Raymond B. Cattell. For Cattell (1957), source traits literally cause the surface consistencies in behavior which are also modulated, to be sure, by situational influences (Cattell, 1963). Moreover, these source traits may be discovered and measured through factor analysis, which identifies dimensions of behavioral covariation that are direct quantitative expressions of traits (Cattell, 1957).

If the circle containing personality structure and trait is dropped from Figure 1 and attention directed instead to the environmental context of behavior, then another realist position more akin to Skinner's (1971) behavioral view is highlighted. Diverse behaviors, including test behaviors, are mutually related because they are elicited and maintained by the same or related environmental conditions, especially reinforcement conditions (cf. Anastasi, 1948; Tryon, 1979). Related behaviors form a response class because they enter the same functional relationships with antecedent, concurrent, and consequent stimulus conditions, and these environmental contingencies control their operation. For trait psychologists test behaviors are largely *signs* of personality structures, whereas for behavioral psychologists they are largely *samples* of response classes. Some psychologists use tests as both signs and samples, being concerned both with predispositions to respond over various environmental conditions *(traits)* and with patterns of reaction under specific environmental conditions *(states)* (Cattell & Scheier, 1961; Spielberger, 1966). In a literal sense, of course, since test behavior is itself behavior, all tests are both signs and samples.

The behaviorist stance, whether realist or phenomenalist, is especially congenial to an operationist approach to measurement, which is not surprising since they both stem from philosophical roots in logical empiricism. Bechtoldt (1959), for instance, claims that "[n]early all postulated constructs listed in the papers on construct validity are . . . of the defined type" (p. 625). As an example, he quotes Lord (1952) on the measurement of ability: "Since, in the final analysis, the only observable variables under consideration are the item responses, any operational definition of ability for present purposes must consist of a statement of relationship between ability and item responses" (p. 4). Bechtoldt (1959) then suggests that "a considerable number, if not all, of the so-called theoretical formulations in psychology dealing with postulated attributes simply use complex definitional transformations of a person's performance as a substitute for an empirical theory. Such transformations . . . are neither theories nor laws . . . [but] provide only definitions of concepts" (p. 625). But Lord's test score model applies to any variable whose item responses display the structural relations specified. It is a scoring or measurement model for relating item response data to test scores, or latent trait parameters, and as such it bears on what Loevinger (1957) called "structural validity."

Such a test score is an intervening variable, in the technical sense that a measure of its amount can be derived directly from empirical laws among the item responses by grouping of terms (MacCorquodale & Meehl, 1948). It becomes a hypothetical construct when interpreted as "verbal ability" or "quantitative reasoning" or "attitude toward war," which is an entirely different matter entailing laws not between item responses but between constructs, each of whose measures has its own measurement model (Messick & Ross, 1962). Bechtoldt's argument identifies not just the meaning of the test score but the meaning of the construct with the scoring operations and thereby confuses the test with the construct and the measurement model with substantive theory.

Inductive Summaries and Surplus Meaning

The interpretation of test behaviors in terms of constructs having surplus meaning beyond the scoring operations but no supposed reality outside the theoretical system of the investigator is portrayed in Figure 2.

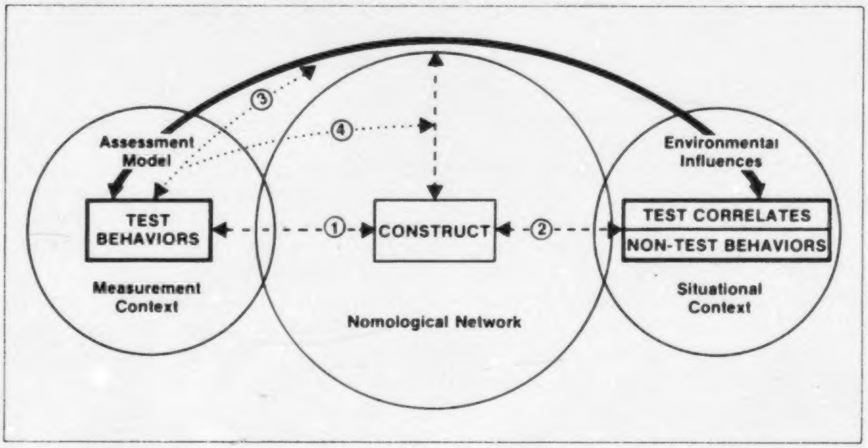


Figure 2

In this viewpoint, "A construct is defined implicitly by the network of associations or propositions in which it occurs," called the nomological network (Cronbach & Meehl, 1955, pp. 299-300). Although proponents of this approach view a construct as "some postulated attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283), strictly speaking they ascribe only systemic existence to it, seeming reluctant to attribute reality to either the constructs or their referents, the postulated attributes or traits.

It should be noted that just as environmental contingencies offered an alternative to traits as sources or determinants of behavioral consistency in the realist worldview, so may postulated attributes of stimuli and situations serve as constructs in this constructivist worldview. An example is Murray's (1938) concept of *press*, which is a directional effect exerted by an object or situation that is classifiable in terms of the kinds of benefits or threats it affords the individual. Similarly, postulated attributes of person x environment interactions also comprise constructs, as in Murray's (1938) concept of *thema*, which is the dynamic structure of a press-need combination. With situational and interactional constructs, Campbell's (1960) principle of "trait validity"—the extent to which a measure relates more highly to different methods for assessing the same construct than it does to measures of different constructs assessed by the same method—still provides an important standard but in this context carries a particularly unfortunate misnomer. Thus, although construct validity is usually invoked in connection with the measurement of traits, it applies with equal force to the measurement of situational and interactional attributes. Our understanding of the functional characteristics of situations in relation to individual differences in human performance is not well developed at this point in either educational or organizational psychology, but important systematizing efforts are underway (Frederiksen, 1972), and this critical measurement challenge animates much of the activity in the burgeoning new field of environmental psychology (Stokols, 1978).

The circle representing the nomological network in Figure 2 is drawn to overlap the situational context as well as the measurement context, to emphasize the need to take account of these factors in construct interpretation, either through a unified construct-context theory or through collateral theories of context effects. The three dashed lines represent three major ways that constructs are inferred—from consistencies in test behaviors, from consistencies in non-test behaviors, and from consistent relationships between the two. The double arrows on those lines indicate that constructs are not only inferred from such consistencies, but in other circumstances they also imply or predict such consistencies.

The numbers in the diagram refer to some recurrent questions in construct validity (Loevinger, 1957, p. 643): (1) "To what extent does the test measure whatever is specified by a given construct?" By and large, this corresponds to "trait" validity. (2) "To what extent does the construct embody a valid hypothesis?" This corresponds to nomological validity. (3) "To what extent does the test measure . . . [something] that 'really' exists?" This is the basic question of behavioral consistency in test performance, which indicates that "something" is indeed being measured. (4) "How well does the proposed interpretation correspond to what is measured by the test?" This refers essentially to substantive validity, which assesses how well the test interpretation captures the nature of that "something." Question 3 examines the magnitude and consistency of relationships, both internal to the test and external to other measures, that is, the structural aspects of internal and external relationships. Question 4 examines the content and nature of internal and external relationships, that is, their substantive aspects.

From this logical-constructivist standpoint, the main function of constructs is to provide inductive summaries of observed relationships as a basis for elaborating networks of theoretical laws (Beck, 1950; Cronbach & Meehl, 1955). Constructs thus provide organized interpretations of observed behaviors as well as a means of predicting previously unobserved behavioral consistencies from the theoretical implications of the nomological network.

Constructions of Reality

A critical-realist approach to the interpretation of test and non-test behaviors is diagrammed in Figure 3.

This perspective is realist because it assumes that traits and other causal entities exist outside the mind; it is critical-realist because it assumes that these entities cannot be perceived directly but must be viewed through constructions of the mind (Cook & Campbell, 1979). Just as constructs were not limited to postulated traits but embraced attributes of situations and interactions, so do these causal entities also include situational and interactional forces. Perhaps a more descriptive label for this position would be "constructive-realist." By attributing reality to causal entities but simultaneously requiring a logical construction of observed relationships, this approach aspires to attain the predictive richness of the realist position while limiting its naive excesses through logical analysis; at the same time, it aspires to retain the summarizing advantages of the constructivist view. As Beck (1950) put it, "The function of constructs is to preserve and summarize; their virtue is metaphysical chastity, not scientific pregnancy" (p. 78).

An especially vigorous proponent of constructive-realism is Jane Loevinger (1957), who argued that the term "*Construct* connotes construction and artifice; yet what is at issue is validity with respect to exactly what the psychologist does not construct: the validity of the test as a measure of traits which exist prior to and independently of the psychologist's act of measuring. It is true that psychologists never know traits directly but only through the glass of their constructs, but the data to be judged are manifestations of traits, not manifestations of constructs" (p. 642). This point is extended here, of course, to include manifestations of situational and interactional pressures as well.

By including both real traits and hypothetical constructs

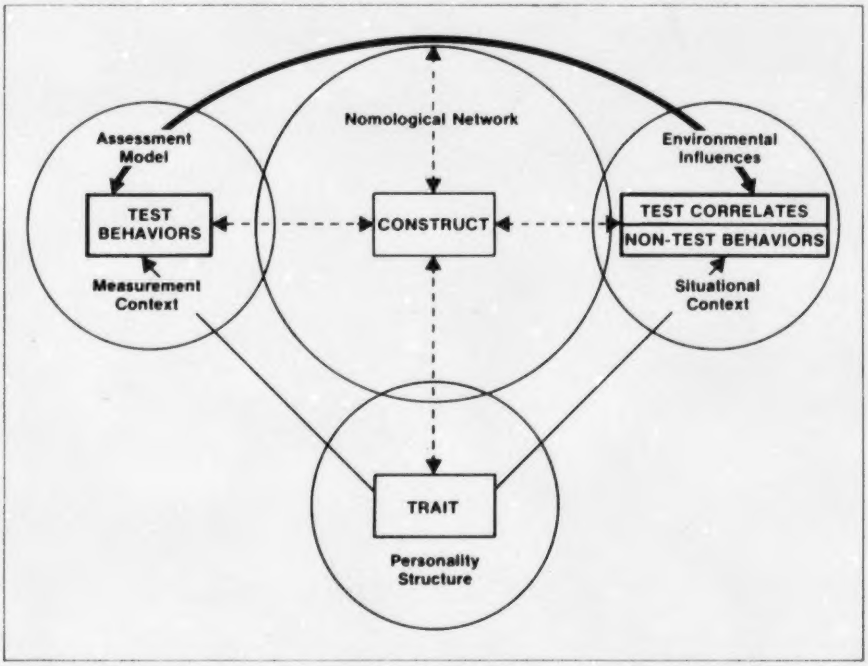


Figure 3

as mediators, this perspective affords a new line of relationship, namely that between the construct and its referent trait. Loevinger (1957) views this relationship as analogous to the one between a statistic and a parameter. "The parameter is what we aim to estimate; the corresponding statistic represents our current best estimate of it. Just so, the trait is what we aim to understand, and the corresponding construct represents our current best understanding of it" (p. 642). This distinction between construct and trait has manifold implications for methodology. Just as a statistic can be biased in its estimation of the parameter, so now can a construct be biased in its representation of the trait. It can be biased, for example, by virtue of its inclusion in a particular theoretical network that is incomplete or tangential to the operation of the trait, or it can be biased by an ideological overlay of value connotations about the nature of man and society. This suggests a possible need to develop alternative trait theories or countertheories to offset the theory-dependence of data collection relevant to the trait, thereby increasing the variety of illuminative data, as well as the need to examine the value implications of relevant construct frameworks and underlying ideologies (Messick, 1975, 1979b).

It would also seem to follow that if traits exist in some sense, then functional characteristics of the person are there to be measured, which puts the primary methodological focus on measurement rather than on prediction or utility (Loevinger, 1957). In this view, response data are first combined not to maximize prediction or utility, but to maximize the meaningfulness of interpretation. The resulting interpretable scores then provide empirical components for prediction systems and rational components for informed decision making. This approach is consistent with the stance of many factor analysts, although factor analysis is not needed to implement it; it implies a simple structure notion of some sort, or at least some notion of structure simple or otherwise, although not necessarily *the* simple structure notion advanced by Thurstone (1947; Tucker, 1955). This is not meant to imply that the generality of factor analysts believe that the generality of factors reflect real traits, but rather that factors are constructs (Royce, 1963) that reflect functional characteristics of performers, or at least of the performance, and that they provide psychologically meaningful independent and dependent variables for scientific and applied prediction. The attribution of these functional consistencies to real traits is explicit in the work of some factor analysts and implicit in the work of many.

If the terms "trait" and "personality structure" are replaced in Figure 3 with terms like "response class" and "stimulus conditions," then the diagram comes to represent the position of many social behaviorists, who do not so much eschew constructs as they do the attribution of traits. This polarity of viewpoints is so widespread that a dimension of individual differences has been posited that contrasts at one extreme individuals who believe their behavior is personally caused and internally controlled with those at the other extreme who believe themselves to be externally controlled pawns of environmental forces (deCharms, 1968; Lefcourt, 1976). Trait theorists treat this belief as a motivational disposition, whereas social learning theorists (Rotter, 1966) view it as a generalized expectancy reflective of past reinforcement history in a variety of situations. Instead of traits, relevant constructs for social behaviorists would be varieties of response classes conceived in terms of the categories of stimulus conditions that elicit and maintain them, along with notions of subjective expectancy of reinforcement and value of reinforcement. If personal qualities are addressed at all, they are conceived not as dispositions but as response capabilities (Wallace, 1966). The central measurement problem is to ascertain whether specific response capabilities exist in the person's repertoire, because task performance is a function of both response capability and the extent to which the criterion situation entails stimulus factors functionally related to the response class (Wiggins, 1973). Thus, tests in this context are usually viewed as *samples* of response classes, and in the extreme, "test scores derived from such behaviors are but descriptions of those behaviors" (Tryon, 1979, p. 405).

This type of behavioral analysis, attempting to identify situational factors that elicit and maintain behavioral consistencies, has much in common with job or task analysis, which attempts to identify those features of the job or task that elicit and maintain adaptive performance, that is, to identify job requirements and the associated classes of behavior that fulfill them. One may go from job analysis to a predictive test, however, in a number of ways, as the discussion thus far makes clear. For example, one might present the identified stimulus or task conditions in roughly the same form and in the same relationships as they appear on the job, thereby eliciting a response class of interrelated behaviors similar to that on the job, although not necessarily the identical class. Such a job-sample test is tailored to a specific job and should be fairly predictive of specific job performance (Wernimont & Campbell, 1968), although prediction may be attenuated by context effects, such as an evaluative atmosphere, that influence the test and not the job, at least not in the same way, or by situational factors, such as interpersonal attraction, that influence the job but not the test. Or one might present the identified stimulus conditions more discretely to cover the major job components in some representative fashion but not in the combinations or interactions characteristic of a specific job. Such a content-sample test may not be as predictive as a job-sample test for specific jobs, but it might have wider applicability across jobs containing various combinations of the task components. Or, as a final instance, one might infer from the categories of task requirements and associated response classes—that is, from what needs to be done and what is done—the processes and abilities suitable for job performance, and then develop or select tests to measure those constructs or traits. This extends behavioral analysis to a deeper level of process analysis as a basis for inferring ability requirements. The resulting construct-referenced test battery may have still wider applicability across jobs and criterion settings calling for those abilities and traits, even though specific criterion task components may be ostensibly different. This treatment of the constructive-realist viewpoint is not meant to imply that for every construct there is a counterpart reality in the person, the situation, or the interaction. On the contrary, many useful constructs, especially higher-order constructs such as "ego" and "self," are employed within this framework as heuristic devices for organizing observed relationships with no necessary presumption of real entities underlying them. The presumption of real entities underlying such constructs as "working class" and "middle class" or "childhood" and "adolescence" is similarly arguable. Rather, the constructive-realist position of Figure 3 might be thought of as an overlay of Figure 2 on Figure 1, where some constructs in Figure 2 have real trait (or situation or interaction) counterparts in Figure 1 while other constructs do not.

In the remainder of this paper, attention will be focussed on that portion of Figure 3 relating test behaviors and their associated assessment models to constructs and their nomological networks, in an effort to highlight some slippery methodological problems in the measurement of educational and psychological constructs. Before leaving these diagrams, however, we should mark the cautionary words of those who remind us what the theoretical edifice is built on. We have sets of observables and the relations between them—the darkened boxes and lines in the three figures—and beyond that, nothing. ". . . [I]n the ultimate analysis, the 'measurement' and 'validation' of constructs can consist of nothing more than the determination of internal structures and cross-structures. . . . If it is useful to refer to these two steps as 'construct validation,' probably no harm is done, but it is important to understand what is being proved and how the related evidence is gathered" (Nunnally, 1967, pp. 97-98). Such cold caution is needed periodically to constrain the conceptual imperialism of personality theorists. It is offered here in the spirit of dialectical openness, on the chance that a confrontation of counterperspectives might illuminate their basic assumptions (Churchman, 1971).

Nomological Looseness in Test Interpretation

In the measurement of educational and psychological constructs, a number of interpretive confusions have arisen about the test in relation to the construct, about the test score in relation to the assessment model, and about the assessment model in relation to the nomological network.

Nomological Noise

One serious but common confusion occurs when a single test is used to represent the construct in studies intended to elaborate the construct's nomological network. Since no such test is a pure exemplar of the construct but contains variance due to other constructs and method contaminants, there is no solid basis for concluding that an observed relationship stems from that part of the test variance reflective of the construct or from other parts. The assimilation of such relationships into the construct's nomological network leads to confusion, which becomes even worse if different single tests are used to represent the construct in different studies. This is the situation with respect to the cognitive style of field dependence versus field independence, for example. Numerous studies use only the embedded-figures test to represent the construct while other studies use only the rod-and-frame test. The variance shared with a third variable by either one of them singly may or may not be the same variance they share with each other (Wachtel, 1972). An especially insidious version of this confusion occurs when the test is presumed to carry the construct into a new relationship, as when the cognitive style of field articulation is viewed as contributing to the formation of a leveling-sharpening factor because a block design test, usually an indicator of field articulation, received a substantial loading on that factor (Santostefano, 1978). Research to sort out and separately score the multiple processes operating on complex tests would improve the situation, but the best way to reduce this nomological noise is to use multiple measures of constructs routinely in network explorations.

Model Slippage

Another type of confusion occurs when scores derived by one assessment model are interpreted in terms of a different assessment model. For example, many of the scales of the Minnesota Multiphasic Personality Inventory (MMPI) were developed empirically by selecting items that significantly discriminated one or another psychopathological group from normals. The assessment model for these scales was a classification system which linked scale scores to the probability of diagnostic placement. Roughly speaking, the higher the score, the higher the probability that the subject belongs in the diagnostic category. But that is not the way the test is typically employed; it is used instead to make inferences about the personality of the respondent. The implicit assessment model for that kind of interpretation is a cumulative scale where the higher the score, the more of some underlying characteristic the respondent displays. Classification studies of the accuracy of patient assignment to diagnostic categories would bear directly on the validity of the scales as originally derived but only indirectly, at best, on the validity of the personality interpretations. The latter assessment model requires scale homogeneity to some degree consistent with the breadth of the trait, and since inferences are usually made about personality differences within both the normal and pathological range, it requires within-group homogeneity to boot. Foremost, of course, score interpretations in terms of personality characteristics require construct validity.

Loevinger (1957), for one, recognized the need to tailor the structural characteristics of the measurement or assessment model to parallel the structural characteristics of non-test manifestations of the trait, so that item responses would be combined into test scores in a trait-consistent fashion. The degree of fidelity achieved was called structural validity, but that principle is contravened if slippage occurs from one assessment model used in score derivation to a different assessment model used in score interpretation. The model underlying interpretation should prevail, since the validity of interpretation is paramount, which suggests that MMPI scales should be appraised more in terms of construct validity and homogeneity than classification accuracy.

Model Compounding

A third type of interpretive confusion occurs not so much from slippage between assessment models as from compounding of assessment models. Sometimes two or more assessment models are combined, or one is overlaid on the other, so that one aspect of test interpretation derives from one model and another aspect from the overlay. This leads to confusion as to which nomological network to reference, or which part of a unified network, as well as to confusion about the kinds of evidence needed for the construct validity of the compound interpretation. As illustrations, we will briefly consider three instances from educational measurement: the assessment of minimum competency, of mastery learning, and of aptitude-treatment interactions.

Consider first an interpretation of test scores in terms of *minimum competency* in a *basic skill* like *reading comprehension*. The primary assessment model, the one underlying score development, is a cumulative scale relating higher test scores to higher amounts of the measured reading competency. Overlaying this is a categorical model relating score levels above a specified standard to the attainment of minimum competency and those below that standard to failure. The example given is further complicated by the inclusion of the term "basic skill" in the interpretation, because that has its own assessment implications. The three components of minimum competency, basic skill, and reading comprehension tap into different nomological networks, leading to confusion as to what construct validity evidence is needed to support the score interpretation. Whatever evidence is adduced usually bears on the interpretation of the scores as measures of reading comprehension. This evidence is typically of the content validity type, although recognition is gradually being given to the need for construct validity evidence for such educational tests, especially evidence to discount plausible rival interpretations of low scores as reflective not of incompetency or reading deficiency but of low motivation, inattention, anxiety, or the like (Messick, 1975, 1979a).

Construct validity evidence for interpreting the scores as basic skill or minimum competency is not nearly as forthcoming, partly because the nomological networks or even the intuitive notions for these constructs are not well elaborated. Is a basic skill one that enables acceptable performance in a wide variety of situations at some functional level of environmental complexity or difficulty? Is it a skill that facilitates later learning in the same area or across areas? Is it a survival skill necessary for societal adaptation? Is it a skill where deficiency carries enduring social stigma? Or is it any one of these, or all of them, or some combination? Evidence bearing on the interpretation of basic skill qua basic skill in any of its possible guises is rare, as is evidence that any particular consensual standard of competency is minimal, although the situation is improving somewhat on the latter score as judgmental and contrasting-group methods for setting standards evolve (Glass, 1978; Hambleton, 1980). It should be emphasized that the concern here is not with social and political issues in the use of minimal competency tests of basic skills, but with the quality of evidence for the validity of the interpretation. When test scores bear the burden of compound interpretations, validity evidence seems to be directed more toward one component than toward others. In this example, the evidence was oriented more toward the primary assessment model and its trait interpretation of scores than toward the overlay models and their educational and sociopolitical interpretations. Some might argue that the concern is misplaced, that "minimum competency" and "basic skills" are part of the rhetoric of test use, not part of the inferences from test scores. There may be a certain truth to that, but it is not likely to be wholly true.

In regard to the measurement of mastery learning in a subject-matter field like science or algebra, a similar overlay of assessment models occurs. A cumulative score model for science knowledge, for example, is overlaid with a categorical score model for mastery states. As in minimum competency testing, at issue is evidence for the validity of a cutting score or standard, in this case for distinguishing mastery from non-mastery, which is essentially a question of the validity of classification assignments to mastery states. Primarily at issue, however, is the *meaning* of mastery as used in test score interpretations. The problem of the standard is usually finessed in mastery learning studies through some form of normative comparison. For example, one widespread approach is to use identical or parallel achievement tests in classes taught by mastery methods and non-mastery methods and "to set the level required for a grade of A in a non-mastery class as the definition of *mastery* for the mastery classes" (Bloom, 1976, p. 4). But when a test score is interpreted as indicating mastery in elementary algebra, for example, what is implied beyond a level of mathematics achievement inferred from the content coverage and construct validity of the test as referenced to the norm of A work in standard classrooms, and what evidence should be adduced to support those implications?

Another kind of overlay of assessment models occurs when a test of spatial ability, say, or need for achievement is interpreted as an aptitude facilitative of learning in a particular educational program. A cumulative score model for measuring the ability or trait is overlaid with a predicted score model for assessing aptitude; that is, individual differences in aptitude are reflected in differences in predicted outcome scores. The same trait measure may thereby function as aptitude to different degrees, or perhaps sometimes even in different ways, for different outcome measures or different treatments. Treatment-interactive aptitudes are those that predict common fixed outcomes differently for one treatment than for another. An additional categorical score model assigning students in different score ranges to different treatments is overlaid at this point on the predicted score model of aptitude overlaying the cumulative score model of the ability or trait.

Two related but distinct notions of aptitude are employed by Cronbach and Snow (1977). Generally, aptitude is considered to be "any characteristic of a person that forecasts his probability of success under a given treatment" (p. 6), that is, aptitude as a *forecaster* of learning or performance. "Psychologically, aptitude is whatever makes a person ready to learn rapidly in a particular situation" (p. 107), that is, aptitude as a *facilitator* of learning. The interpretation of scores on an ability or trait measure as indicative of a treatment-interactive aptitude is frequently buttressed by construct validity evidence for the trait measure and by predictive evidence of aptitude not only as a forecaster of learning but as a differential forecaster for different treatments, but rarely is evidence forthcoming for aptitude as a facilitator of learning. Validity of this latter implication in the form of evidence illuminating the process of facilitation is particularly important if we intend "to make the development of aptitude a deliberate goal of instruction" (Cronbach & Snow, 1977, p. 521).

These compound interpretations of test scores encumber the construct validation process with the need for multiple evidence bearing on different nomological connections. Since the overlay construct theories are at least one step removed from the primary theory underlying test development, it is not always readily apparent which features of the overlay construct should be emphasized in the construct validation of the compound interpretation. Some help in this regard might be derived from consideration of the associated assessment models involved, since they may contain within them tacit signals about aspects of the constructs that are salient for score interpretation. This notion is essentially an obverse of Loevinger's (1957) principle of structural fidelity, for if an assessment model should be chosen to reflect the structural relationships in non-test manifestations of the trait, then the assessment model used should imply a set of structural trait characteristics that are consonant with it, or at least a variety of characteristics that are reasonably compatible with it.

Assessment Models As Implicit Psychological Theories

Since we consider assessment models broadly to include both the measurement or score model and the conditions of measurement, the important implications for our purposes focus upon the structure of individual differences resulting from the compounding of a trait measure with the conditions of testing for minimum competency or mastery learning or aptitude-treatment interaction.

In minimum competency testing, the distribution of individual differences along the cumulative scale of the skill measure is cut off at the minimum competency standard. Individual differences below the standard are eliminated by remedial work while individual differences above the standard, strictly speaking, are unspecified. One source of confusion in the interpretation of minimum competency derives from the surplus meaning associated with the educational programs in which the concept is embedded, particularly in regard to the extent to which those programs are mute concerning individual differences above the standard or specifically attend to skill development throughout the range. Evidence about changes in the distribution of individual differences before and after or during the course of minimum competency programs would not only clarify the nature of those programs, but the nature of the construct and the meaning of the standard in terms of the value implications of the distributional changes (Glass, 1978).

In testing for mastery learning, the distribution of individual differences along the cumulative scale of the achievement measure becomes systematically compressed toward the upper end, and in the extreme of all students achieving mastery of the given educational goals, individual differences in outcome would be eliminated. One possibility, of course, is that individual differences do not disappear but are converted into individual differences in rate of learning. But several mastery learning programs report relatively small variations in learning time, with about four-fifths of the students reaching a level of achievement which less than one-fifth attain under non-mastery conditions in only about 10 to 20 percent additional time over scheduled classroom time (Block, 1971, 1974; Block & Burns, 1977; Bloom, 1976). Another possibility is that individual differences in outcome do not disappear but are merely obscured by examining only the fixed objectives of the given course. As Greeno (1977) put it, "not all students acquire the same knowledge, even when they all pass the same test. Some of the students achieve the specific instructional objectives and acquire more knowledge in the process" (p. 105). Evidence about the extent of individual differences in depth of knowledge and understanding among students all of whom have mastered the fixed curriculum objectives would help clarify the nature of mastery learning as a construct, especially in its relation to the concept of equality of learning outcomes.

In the assessment of treatment-interactive aptitudes, there is also an implication that individual differences in outcome will be reduced by placing students into different treatments tailored to their functional characteristics as learners. There is a further implication that the same fixed goals of achievement can be reached by the different instructional methods. But as with mastery learning, students achieving the specific educational goals may still differ in what else they acquire, and since the treatments differ in procedure, these differential learnings may entail process more than content. Thus, as Carroll (1967) pointed out, "The achievement of fixed, common goals may be attained, if this is true, only at the cost of *differential* achievement of other goals" (p. 43). Evidence about the range and nature of individual differences in treatment-related cognitive processes and in depth of knowledge and understanding among students achieving the same fixed goals by different methods would clarify the meaning of treatment-interactive aptitude as a construct and its relation to other educational constructs (Snow, 1977).

Overview

To recapitulate briefly, there are different points of view about the import of test behaviors in relation to non-test behaviors and about the role of mediating theories as guides to empirical reality. The use of constructs to develop intuitive systems of meaning for observed relationships appears to be a fruitful heuristic if buttressed by convergent and discriminant operations to validate the interpretation of test scores as measures of constructs and to validate observed relationships as reflective of meaningful connections in the nomological system. Furthermore, if test and non-test behaviors are viewed as manifestations of real traits, albeit perceived through the categorical filters of constructs, then predictive richness beyond the limits of a particular construct theory might accrue, with the added possibility of contrasting alternative or conflicting construct theories of the trait and of probing possible biases in constructs as representations of traits. In addition, each construct measure has an associated assessment model which should be coordinated with the construct theory, and any looseness in that coordination leads to interpretive difficulties.

Against this background, constructs and construct validity have been seen to have their ups and downs in educational and psychological measurement. On balance, I think in psychological measurement there were more ups and in educational measurement, more downs.

References

Anastasi, A. The nature of psychological 'traits.' *Psychological Review*, 1948, 55, 127-138.

Bechtoldt, H. P. Construct validity: A critique. *American Psychologist*, 1959, *14*, 619-629.

Beck, L. W. Constructions and inferred entities. *Philosophy of Science*, 1950, *17*, 74-86.

Block, J. H. (Ed.). *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston, 1971.

Block, J. H. (Ed.). *Schools, society, and mastery learning*. New York: Holt, Rinehart and Winston, 1974.

Block, J. H., & Burns, R. B. Mastery learning. In L. S. Shulman (Ed.), *Review of research in education* (Vol. 4). Itasca, Ill.: F. E. Peacock, 1977.

Bloom, B. S. *Human characteristics and school learning*. New York: McGraw-Hill, 1976.

Bowers, K. S. Situationism in psychology: An analysis and critique. *Psychological Review*, 1973, *80*, 307-336.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 1960, *15*, 546-553.

Carroll, J. B. Instructional methods and individual differences. In R. M. Gagne (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles E. Merrill, 1967.

Cattell, R. B. *Personality and motivation structure and measurement*. Yonkers, N. Y.: World Book, 1957.

Cattell, R. B. Personality, role, mood, and situation-perception: A unifying theory of modulators. *Psychological Review*, 1963, *70*, 1-18.

Cattell, R. B., & Scheier, I. H. *The meaning and measurement of neuroticism and anxiety*. New York: Ronald, 1961.

Churchman, C. W. *The design of inquiring systems: Basic concepts of systems and organization*. New York: Basic Books, 1971.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281-302.

Cronbach, L. J., & Snow, R. E. *Aptitudes and instructional methods*. New York: Wiley, 1977.

deCharms, R. *Personal causation*. New York: Academic Press, 1968.

Ekehammar, B. Interactionism in personality from a historical perspective. *Psychological Bulletin*, 1974, *81*, 1026-1048.

Frederiksen, N. Toward a taxonomy of situations. *American Psychologist*, 1972, *27*, 114-123.

Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, *15*, 237-261.

Greeno, J. G. A review of *Human characteristics and school learning*. *Proceedings of the National Academy of Education*, 1977, *4*, 85-116.

Hambleton, R. K. Test score validity and standard setting methods. In R. Beck (Ed.), *Criterion-referenced measurement: State of the art*. Baltimore, Md.: Johns Hopkins Press, 1980.

Jackson, D. N. The dynamics of structured personality tests: 1971. *Psychological Review*, 1971, *78*, 229-248.

Lefcourt, H. *Locus of control: Current trends in theory and research*. New York: Halstead Press, 1976.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, *3*, 635-694 (Monograph Supplement 9).

Lord, F. M. Theory of test scores. *Psychometric Monographs*, 1952, *72*, Whole No. 7.

MacCorquodale, K., & Meehl, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 1948, *55*, 95-107.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Messick, S. Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, 1979, *71*, 281-292. (a)

Messick, S. Test validity and the ethics of assessment. Invited address presented at the meeting of the American Psychological Association, New York, September, 1979. Manuscript submitted for publication, 1979. (b)

Messick, S., & Ross, J. Psychological structure and measurement models in personality assessment. In S. Messick & J. Ross (Eds.), *Measurement in personality and cognition*. New York: Wiley, 1962.

Murray, H. A. *Explorations in personality*. New York: Oxford University Press, 1938.

Nunnally, J. *Psychometric theory*. New York: McGraw-Hill, 1967.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, *80* (1, Whole No. 609).

Royce, J. R. Factors as theoretical constructs. *American Psychologist*, 1963, *18*, 522-528.

Santostefano, S. *A bio-developmental approach to clinical child psychology: Cognitive controls and cognitive control therapy*. New York: Wiley, 1978.

Skinner, B. F. *Beyond freedom and dignity*. New York: Knopf, 1971.

Snow, R. E. Research on aptitude for learning. In L. S. Shulman (Ed.), *Review of research in education* (Vol. 4). Itasca, Ill.: F. E. Peacock, 1977.

Spielberger, C. D. Theory and research on anxiety. In C. D. Spielberger (Ed.), *Anxiety and behavior*. New York: Academic Press, 1966.

Stokols, D. Environmental psychology. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology*. Palo Alto, CA.: Annual Reviews, 1978.

Thurstone, L. L. *Multiple-factor analysis*. Chicago, Ill.: University of Chicago Press, 1947.

Tryon, W. W. The test-trait fallacy. *American Psychologist*, 1979, *34*, 402-406.

Tucker, L. R. The objective definition of simple structure in linear factor analysis. *Psychometrika*, 1955, *20*, 209-225.

Wallace, J. An abilities conception of personality: Some implications for personality measurement. *American Psychologist*, 1966, *21*, 132-138.

Wachtel, P. L. Field dependence and psychological differentiation: Reexamination. *Perceptual and Motor Skills*, 1972, *35*, 179-189.

Wernimont, P. F., & Campbell, J. P. Signs, samples, and criteria. *Journal of Applied Psychology*, 1968, *52*, 372-376.

Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.

DISCUSSION

Glymour: I don't understand what you mean as between the naive viewpoint, Figure 1, and the critical-realistic in Figure 3. I mean, is this critical-realistic view just a view; what's really going on is these traits but we might be wrong about them?

Messick: The difference between the conception of the trait and the realist view is—the only reason I made a distinction is that when I argue with some colleagues, it's apparent to me that the concept of having some kind of logical analysis of the reality of the trait relationships they are talking about is not a necessity for them. So it's adding to the realist's view the notion of some kind of logical construction that would limit some methodological excesses.

Glymour: Let me press you a bit more. What is it you mean by logical construction? What is it you are logically constructing in the first place?

Messick: That would be the observed relationships, in the sense of following through the nomological connection of theoretical relationships, rather than a vague statement that the trait is operating and, if the expected result isn't obtained, it's because of some other trait interference, through the personality structure. That trait interference might become nomological, but it's not enough to say another trait is operating, and that's why this result didn't look sensible.

Glymour: Even a naive realist like me believes that you ought to have explicit, clear theorizing, and it doesn't come to more than that.

Messick: When I started out, I didn't think it was necessary to have anything except Figure 3, until I argued with some people who are really adhering to Figure 2, who believe in neither the reality of traits nor the reality of anything else except the observables.

Glymour: The difference between Figures 1 and 2 is clear.

Sternberg: I want to expand on your answer a little; it seems so germane to the point you were raising yesterday.

As I understand the difference you were making, you were saying, 'Let's take it in the context of intelligence'. If you have a trait of intelligence which you define as an ability to adapt to the environmental circumstances of your world, that would be your basic trait and you would come up with some constructs that measure that—say the existing psychometric tests.

Now it may be that you then come up eventually with some better set of tests; and what changes then, as I understand it, is this construct of the nomological network, but you don't then have to go back and change the basic trait definitions of ability to adapt, but, rather, you have a better nomological network and system for measurement. If you don't make that distinction, you don't have a clear distinction between the operations used for measuring the trait in terms of the construct and your underlying idea of adaptiveness, which could be measured in several ways.

I see the distinction as necessary and, moreover, as being consistent with what you were saying yesterday in your talk.

Glymour: I quite agree one does have to keep those matters clear, and simply because you find a new way to determine atomic weights doesn't mean you are talking about something different because it was determined otherwise.

Cronbach: I must say at least two things. First, I think we would agree completely with what you were saying about the network, and it permits me to emphasize something I think is not adequately emphasized, and that is that we are construing the situational context at the same time that we are construing the personal. It is only historical accident that construct validation got associated with trait considerations. We have to divide up the situational world just as much as the treatment world, if you will, and this means the employment world also. Jobs are, in this sense, constructs; they are ways of construing what people do when we put them into categories.

We got a nice example earlier of the process of construing the situation when Bob Sternberg was talking. First there was the treatment, there was an induction of a figural-modal approach to certain problems and then, in the course of his analysis, he construed it differently by bringing in the notion that some people were following the induction and some were not, so the relevant variable at this point was the response variable rather than the treatment variable. And this is very much the way Snow and I would want to think about this kind of problem.

Second, I want also to bring into the record here a mention of the Cook-Campbell book *Quasi-experimentation* using construct validation in connection with treatments, in connection with validation studies. I'm not entirely happy with some of the distinctions they were making there, but fundamentally they are saying exactly this: if we want to draw conclusions about, let's say, a form of management development, we have to think very hard about that form. If we are going to generalize, it's one thing to say it's been factored but it's quite another to put labels on it and generalize and say, Here is a style of management development. So I think it's quite important that anybody who wants to think about construct validation read that book.

Williams: I think you ought to entertain a question from those who are in the world of work, less so in the world of education, as to how this formulation is responsive to some of the kinds of concerns that Eric Prien and others were referring to yesterday.

Messick: As I remember that discussion, there was some question as to whether construct validity would be relevant to the kinds of constructs that would exist in the world of work, constructs, as Lee pointed out, that are derived from job analysis. The answer would be yes, it is relevant. But construct validity as we have been talking about it is essentially evidence to support the validity of the interpretation of the measure. When you get to the use of the measure, in some applied setting, then other things are involved. I think one of the confusions has been to separate use from interpretation to such a degree that there is a feeling that one can get away in applied settings with empirical predictive validities that are never referenced to the interpretation of the test; if you have measures that predict the criterion behaviors well, then that's all you need. But one of the messages from construct validity is that it isn't all you need.

In addition, you ought to consider what the test means because not all predictors are acceptable in applied work. Some of them would lead you to question the meaning of the criterion, for example, so you might revise the way the criterion is being handled. For example, predictors such as biographical scales might implicate sexist and racist characteristics. If you don't understand the meaning of the predictor, you might make the wrong decisions. Even if you do understand the meaning of the predictor, that doesn't mean it's useful in the applied setting. So, in addition to construct validity, you need some estimate of the relevance of the predictive measure and some estimate of the utility of the measure and some estimate of the social consequences of its use before making the applied decision.

Anastasi: Isn't it a fact that construct validity enables you to give a more meaningful interpretation of what's going on and so increases the generalizability of the use of the test in a job situation? Just straight empirical prediction is more difficult to apply to other situations. Doesn't construct validity contribute to that purpose?

Messick: Yes, it does; at least there is that possibility.

DuBois: I would like to make a couple of remarks about a nominal scale. The categories in the nominal scale have to be coordinate and mutually exclusive and comprehensive. Now, the committee that came out with the three types of validity, content, criterion-related and construct—I don't think they had coordinate categories. It seems to me that the way in which the validities have to be put together are entirely in terms of the type of criteria. In other words, all our types of validity are criterion-related, so the term "criterion-related validity" rather makes the issue not quite so clear as I would like to have it. For a defined domain, such as an occupation, construct validity is very important, but there the domain of the criteria is indefinitely large—real-life criteria in real-life situations. However, for content validity the predictors and the criteria tend to be in the same domain more or less identically.

It seems to me that part of the confusion about construct validity has come from lack of clearness in defining our criteria. And, of course, I think we are all enthusiastic about construct validity because the other two types are on the engineering side, whereas construct validity is much more general, much more seeking general truths.

Novick: The discussion that occurred today, put together with what we heard yesterday, leads me to suggest that we at least think about the following idea—and I present this just in these terms to think about it, to see whether or not it will work. And I will say I claim no originality for this.

The Uniform Guidelines present us with three strategies for validating the test, and the way you read it is you can do A or you can do B or you can do C. And it's very clear that from their point of view, A is preferred to B, and B is strongly preferred to C.

What flows out of the discussion here, at least tentatively, is the idea that a good validation study is not A or B or C, but some combination of A and B and C, the particular combination depending upon circumstances, the degree of generalization that is required, the possibility of validation in the circumstances, the money that's available, and the knowledge that's available. I'm not sure that that's a viable general approach to validation, but I think we ought to think about it. But how one takes all of that and somehow gets it into a document like the Guidelines, how one gets some ideas like that into APA Standards, is something we ought to be thinking about.

Messick: It seems to me that the leverage we have to try to do something like that comes from the rational framework that can be provided from the meaningful interpretation of the criterion and the situational setting. Without that kind of rational framework, it would be pretty hard to do in the absence of studies that combine A, B and C.

Would Bob Guion like to elaborate?

Guion: We start with the selection problem, do a job analysis, and from this we draw some inferences about the variables we might want to assess in selecting people. We have a logical problem before we do any measurement of those variables. That is the problem of asking whether these variables are or are not related to performance on the job, conceptually. If we cannot come up with any logical argument for saying that the variable is related to the job, the whole idea goes into the wastebasket.

If we say yes, we do have a logical relationship between the variable and the job, then we have to figure out how to measure that variable. And it is only at this point that the questions of validity of measurement come into play.

We have perhaps some question of the validity of the constructs, in trying to determine whether they are going to be relevant or not. But we usually don't think about that very much because in selection problems we usually take variables that are not sufficiently elegant to worry in traditional ways about the validity of the construct.

And I think much of the confusion comes about because there are two separate questions at this point that have to be asked. One is the question of the validity of the assessment of the variable, and the other is the question of the usefulness of assessing that variable in that selection situation. In our traditional criterion-related validation studies, if we make the assumption that the criterion measure is useful for the job, we are combining these two questions.

When we move into construct validity, or, if you insist, content validity—people do keep using that term in spite of us—then we are talking solely about the validity of the measurement of the assessment, and we are not addressing, at that point, the question of usefulness.

I don't know whether that's really responding to Mel or not, but I think it is, because if we do say, in a new set of guidelines, or in new standards, that the assessment of the validity of measurement is a combination of concern for how the measurement is constructed—or the content validity argument—and the evidences of construct validity—which are necessarily in criterion-related studies—and, therefore, combine all three, we still will not have directly addressed the question of the usefulness of those valid measuring instruments.

I can develop a very valid instrument for measuring spelling ability. All I have to do is come here to Princeton and find some of your files, and I'm sure I can come up with a valid test for spelling. We can identify its content validity in terms of the words chosen and their relation to some standard frequency index. We can identify the construct validity by finding out whether people who are generally identified as spelling error-free in their typing of manuscripts are scoring high on the test, and many other things. But that test is of absolutely no value in the selection of people who don't have to spell anything on their jobs.

Glymour: It seemed to me Mel Novick's proposal was in fact quite right, if we think about the description of the validation procedures for the PACE examination.

The point I want to make is that what is at stake is the confirmation of a kind of theoretical claim, namely, that an instrument is a good predictor. It seems to me that you have to resist the demand that there be rules, an explicit and codifiable kind, as to when such a claim has been confirmed. You are talking about the weight of scientific evidence, and that simply doesn't follow any single procedure from a scientific point of view.

Novick: I would agree with that, but there is the practical problem of the use of tests in education and industry, and of the requirement to satisfy certain regulations, both EEOC and others. So one is somehow forced to try to go a little further than one would in a purely scientific discourse, which is where it becomes extremely difficult. And I don't have the answers to that.

Prien: To respond to something Bob said with reference to the difference between criterion-related and content validity studies, the suggestion he makes is that, with criterion-related validity, there is a criterion problem, and that the two concerns overlap, particularly with reference to utility, which doesn't exist with content validity.

I would suggest that if you analyze the operations, you simply have, on the one hand, a data collection operation in criterion-related validity, and in content validity you substitute a judgmental operation made with reference to either an assumed or hypothesized criterion. There is a difference in the operation but the two are there.

Take a look at rating scales, Bob, that you use for evaluation of job elements by subject matter experts—and that is a criterion-related operation. It's a lot weaker; nonetheless, it exists.

BARBARA LERNER

LEGAL ISSUES IN CONSTRUCT VALIDITY

The elegance of its initial formulation notwithstanding (Note 1), construct validity is only the special name coined by one small subgroup of social scientists to refer to the method by which all scientific theory is developed. Because of that, the questions I have been asked to address here today—questions about legal issues in construct validity— have implications that go far beyond the field of psychometrics. They are, inevitably, questions about the proper relationship between science and law in a democracy.

Truth in Science and in Law

Science is one way of searching for truth in the western sense (Note 2), that is, for coherent, stable, hence predictable relationships between any of the things that contribute to what William James called "the blooming, buzzing confusion" of the universe (Note 3). Theology and metaphysics are older and, in relation to some problems, still arguably better ways of seeking truth. What is special about the scientific method is not any general superiority but a unique insistence upon a single method of deciding whether any idea about the relationship between two or more things is true or not. This method involves subjecting ideas to a series of tests in which the possibility of being proven wrong by empirical data—those "stubborn" (Note 4) Jamesian facts—is kept constantly present.

In psychometrics, as in all other scientific specialties and subspecialties, that requires us to state our ideas as precisely as possible, not in the form of closed logical syllogisms, armored against the rude shocks of reality, but in the form of open, vulnerable hypotheses, specific bets made in advance of the game, and in a manner amenable to statistical testing. Only ideas which have survived this singular test—Karl Popper called it "falsifiability" (Note 5)—may be regarded as truths, or at least, as scientific ones.

Given such a system, it follows that disputes about the truth of competing hypotheses in science can be resolved, if at all, in one way only: by the accumulation of a weight of evidence sufficient to convince all who choose to play the game of science by the rules defining it that one hypothesis is probably true and the others false. When a dispute is over a relatively simple specific question like the existence of differential validity, it can often be resolved fairly rapidly by a series of obvious empirical tests. When a dispute is over an extremely complex general question, it may take many decades or even centuries before it can be resolved in a truly scientific fashion. An example here would be the dispute over whether social science will make greater progress by focusing only on observable behavior or by testing hypotheses about relationships between phenomena that cannot be directly observed.

In the latter case, as in the former, there is nothing for scientists to do but work and wait, withholding final judgment until a sufficient body of empirical data has been amassed, no matter how long that takes. Scientific disputes can never be legitimately resolved in any other way, least of all by authoritative pronouncements. In the scientific realm, evidence alone is decisive.

The scientific definition of truth presented above stands in sharp contrast to the legal one. A legal fact, after all, is technically defined as anything that a duly authorized trier of fact—judge or jury—reasonably finds to be so (Note 6). Judges and juries must make their findings of fact at the conclusion of each trial and they must make them only on the basis of the evidence produced by the parties to the litigation before them, regardless of its adequacy. They cannot defer judgment and they cannot seek additional enlightment by searching for evidence beyond that which the litigants choose to present. And, except under extraordinary circumstances, their judgments on matters of fact are final. Appellate courts exist to review conclusions of law; conclusions of fact are subject to review only when they are patently unreasonable on the basis of the evidence presented.

Now, obviously, this definition of fact or truth does not derive its ultimate justification from evidence alone, still less from a priori evidence alone. Its ultimate justification is derived from authority, backed up by the coercive power of the state. The power of the state and its courts is an awesome power but, in the Anglo-American system, it is made less so by a host of procedural protections developed out of centuries of experience. These procedural safeguards reduce the arbitrariness inherent in any system of human authority but they do not and cannot eliminate it. No system that gives such finality to the judgments of a particular individual or a particular group of 12 individuals at a particular point in time can ever hope to free itself from the dangers of subjectivity and error. It is, however, a great mistake to conclude that law is therefore an inferior method of seeking truth. Such a conclusion rests upon a fundamental misapprehension of the essential nature of law and of its role in human society.

In essence, law is not a method of seeking truth at all; it is a way of seeking other ends, variously defined, and its relation to truth in general and to scientific truth in particular is a very complex one. Truth is the supreme value in science, and the only one. It is not and cannot be the supreme value in law because there it is only a means to the realization of other ends. In the Platonic realm of the ideal (Note 7), these other ends might be summarized by the word justice, but in the real world, justice without order is a contradiction in terms. Accordingly, law has to do with order and with ways of distributing and limiting the power to create and maintain order by resolving potentially disruptive disputes through the formal exercise of authority.

Potentially disruptive disputes must be resolved, and rapidly, lest their potential become actual, threatening the public order or, in the standard legal phrase, the public health, safety or welfare. In most instances, such disputes can only be resolved through the exercise of authority, because the a priori empirical data available at the point when judgment must be rendered are rarely adequate to establish truth in the stringent scientific sense defined above; also because truth, or at least a reasonable nonscientific approximation of it, is a necessary basis for justice, but hardly a sufficient one. Justice, after all, has to do with what ought to be and truth, in the scientific sense, tells us only what is, or at best, what can be. There is no purely empirical way of determining what ought to be. Hence, the necessary use of legal authority, backed up by force when necessary, to resolve the sorts of disputes that must be resolved.

Science, by definition, rejects such authority as a means of resolving disputes within its own limited realm and, as a consequence, scientists cannot allow their disputes to be settled by judicial fiat, lest they cease to be scientists at all. The resultant tension between science and law is inevitable, but open clashes between the two are not. Clearly, the potential for such clashes is ever present and has been actualized on a number of melancholy occasions in the relatively brief history of science (Note 8). However, in western democracies, it has generally been seen as both possible and desirable to avoid such clashes through the exercise of restraint on both sides.

Activism, Authority, and Restraint

Traditionally, Anglo-American jurists have exercised restraint by deferring to the consensus of scientific opinion on purely scientific questions. Responsible scientists have exercised a complementary form of restraint by making the clearest distinctions possible between their scientific judgments and their social policy preferences. These distinctions are sometimes subtle, complex and difficult to draw, but they are crucial to the maintenance of the social and legal climate that has nurtured the growth of both science and democracy in this country for some 200 years.

Scientists, after all, have a right to expect deference from courts only in relation to their scientific judgments. We have no right whatsoever to any form of special, elite treatment for our social policy preferences and any claim to the contrary is subversive to the fundamental notion of equal justice under law. In our legal system, courts are obligated to defer only to policy judgments that reflect the will of a majority of all the people as expressed in legislation enacted by their elected representatives. Courts are, in fact, legally empowered to reject such policy choices and the legislation embodying them only when they may fairly be judged to violate the limits set by the basic document that binds us together, the Constitution of the United States (Note 9).

At any rate, that is how the system is supposed to work, but, in recent years, the traditional system has been under intense and increasing attack by liberal intellectuals in law and related fields. Judicial restraint, particularly, has come to be seen by some of my fellow lawyers as an immoral bias in favor of an unjust status quo, and it has undergone some very significant erosions. The result has been an unprecedented expansion of judicial and bureaucratic activism, powered by something akin to missionary zeal in the name of social justice and equality (Note 10).

Activism on the part of nonelected officials is always problematic in a democracy; it is especially so when it is exercised on behalf of an elite concept of justice, as it is today. The concept of justice that currently inspires much of the federal judiciary and the federal bureaucracy is one that can only be characterized in those terms because it is a concept that has been rejected by a large majority of ordinary Americans of all races and by the legislation enacted by their elected representatives.

Public opinion poll data are very clear and consistent on this point: most Americans, black and white, endorse a concept of justice that requires equal opportunity for all individuals now. They reject the competing idea that justice requires equal results for all groups now (Note 11). This majority concept is written into Title VII of the Civil Rights Act of 1964 (Note 12) with unusual clarity. As the legislative history of that bill reveals, the legislators who passed it understood that there were two competing concepts and they made a very deliberate and explicit choice between them, opting for equal opportunity now and rejecting equal results now (Note 13). If you haven't already looked at the legislation itself—it doesn't take long—look particularly at Sections 703(A) and (D). That policy choice strikes many federal judges and bureaucrats as a seriously inadequate response to the problems of black Americans and those of several other previously victimized groups. Accordingly, they reject it as immoral or insufficiently moral, in favor of the competing policy choice: equal results now.

Thus, we have a situation in which many federal judges and bureaucrats are eager to implement a policy which they have no legal mandate to implement. In such a situation, open implementation is a risky course, and, as a result, the temptation to implement the elite policy preference surreptitiously through disguised and indirect means is strong. During the last decade, some federal judges and many agency officials have succumbed to that temptation (Note 14) in ways that have profoundly troubling implications for the relationship between one group of scientists—psychometricians—and the law.

Simply stated, the problem is this: valid tests are no obstacle to the goal of equal opportunity for all individuals now but they do constitute a significant barrier to the goal of equal results for all groups now. The reason for this is apparent to anyone who is reasonably familiar with the data on test results in different groups: in this country at the present time, the scores of black Americans as a group tend to be lower than those of white Americans as a group on many valid academic and vocational tests (Note 15).

Because of this, there is great pressure to rewrite scientific standards for test validation in accord with the elite policy preference. If this pressure did not exist, the four federal agencies that have issued guidelines on employee selection procedures (Note 16) could have dealt with test validation procedures with a single, obvious sentence: any test validation procedure that satisfies the *Standards for Educational and Psychological Tests* issued by the American Psychological Association and its sister organizations satisfies these guidelines. The federal bureaucracies have rejected this simple, obvious course, choosing instead to try to impose their own definitions of validity on the scientific community, including a particularly unsatisfactory definition of construct validity (Note 17).

Government Guidelines in the 70's

The bureaucratic definition of construct validity is one that is biased by a scientifically unjustified preference that appears again and again, not only in section 14 on construct validity, but throughout the guidelines: a preference for a focus on behavior that can be directly observed and a corresponding hostility to approaches that focus on behavior that cannot be directly observed (Note 18).

Now, most agency bureaucrats, with the single, notable exception of those who have graced the staff of the Office of Personnel Management, are not psychometricians or scientists of any sort and, as a result, few of them care very much one way or the other about the long-standing dispute between behaviorists and non-behaviorists. However, many bureaucrats feel—perhaps not altogether without justification—that they can more easily advance their social policy preferences by espousing the behaviorist position than by espousing the non-behaviorist position. Hence, the behaviorist bias of the guidelines.

Recently, some psychometricians have criticized this behaviorist bias and urged their fellow scientists to reject the guidelines on the grounds that behaviorism is obsolete because it is an outgrowth of logical positivism, a school of philosophy that most contemporary philosophers now reject. With these colleagues, I must respectfully disagree, not on philosophical grounds but on scientific ones. As scientists, I think we are all obligated to adhere to our rule of evidence in deciding what is scientifically acceptable and what is not, and the empirical evidence is simply not sufficient to justify any official dethronement of the behaviorist approach to psychometrics or to any other major area of psychology that I am familiar with.

For whatever it is worth, my own philosophical preference is for a non-behaviorist approach to social science in general and to psychometrics in particular (Note 19), but I would be every bit as appalled to see that bias incorporated into the guidelines as I am to see the opposite bias incorporated and made official there and for exactly the same reason: we do not have enough evidence to warrant the official enthronement of either approach. Unless and until we do, the rules of science require us to acknowledge that both approaches are equally acceptable, and to reject any government attempt to make one approach superior to the other by force of law.

The 1978 guidelines also incorporate a scientifically unacceptable position on another major issue that is relevant to construct validity but hardly unique to it: the question of differential predictive test validities for members of different racial and ethnic groups (Note 20). Bias here is a question of timing, as it often is in science: hypotheses that seem plausible on the basis of early findings frequently become less so later, when more careful analyses are completed and when more and better data have been accumulated.

That, of course, is what happened to the hypothesis that most tests in common use have greater validity for members of some racial and ethnic groups than for members of others. In 1974 when the last edition of the Standards was published, that hypothesis still seemed plausible to many competent psychometricians. As a result, the 1974 Standards required the collection of data relevant to it (Note 21). Although the Standards have no legal force, psychometricians responded for the most part by carefully reviewing and reanalyzing previous research on this question and by conducting some well-designed new studies (Note 22). Reanalyses of the old studies and data from the new studies repeatedly pointed to two conclusions: 1) differential validity, to the extent that it exists at all, is a trivial phenomenon; and 2) tests are no more likely to underpredict minority performance than they are to overpredict it (Note 23).

It is, of course, impossible to specify the precise moment at which enough data had been amassed to convince a substantial majority of all competent psychometricians that the weight of evidence on the question of differential validity pointed to the conclusions specified above. It is, however, quite clear that this was the case by the time the new uniform guidelines were issued in August of 1978. As a result, it was equally clear that the requirement for the continued collection of data on differential validity incorporated into the 1978 guidelines could no longer be justified in scientific terms.

It is there, nonetheless, and it proved very useful to those bureaucrats who wanted to attack a major new construct valid test—the PACE test—as an obstacle to equal results now (Note 24). These bureaucrats found it politic to attack PACE by claiming that it was not adequately validated. They based their claim on the fact that data sufficient to disprove, once again, the differential validity hypothesis were not immediately available for this test (Note 25). And, regrettably, they could point to the 1974 Standards to support their stance on this issue.

Scientific Standards for the 80's

The new Standards, scheduled to appear in 1981, will, I trust, correct the discrepancy between the current scientific consensus and the published statement of it. At the very least, the accumulated empirical evidence against the differential validity hypothesis is strong enough to shift the scientific burden of proof to the shoulders of those who still regard that hypothesis as plausible.

My trust that the new Standards Committee will deal with this issue objectively as scientists and not subjectively as social policy advocates is based upon my belief that most psychometricians remain firm in their commitment to respect scientific evidence, regardless of the diversity of their personal preferences concerning definitions of justice and equality and the contrasting social policy choices that emerge from them. It would not, however, be prudent to be overconfident about this matter. In science as in law, traditional standards are under intense and increasing attack today.

As I noted in a recent paper (Note 26), those who have grown cynical and suspicious about almost everything in American life often see those standards as a mere cloak for subjective political preferences, an effort to resist needed change by hiding behind a mask of pseudo-objectivity. From this perspective, the only acceptable research results are those that confirm prior assumptions about sin and salvation in American society, and those who produce evidence contrary to such assumptions are viewed as "insensitive" at best, as racists, Toms and reactionaries at worst.

In fact, however, much of the evidence against those assumptions has been produced by scientists who shared them, and who reported their contrary findings with honest distress, in this area as in a number of other, related ones (Note 27). To these colleagues, particularly, I would like to emphasize the fact that rejection of the differential validity hypothesis is not tantamount to rejection of the equal results now concept of social justice.

As I have indicated in previous papers (Note 28), I personally reject that concept, but scientists who accept it can do so without any loss of scientific integrity as long as they do not distort clear empirical evidence in order to pretend that valid tests are invalid when their results disconfirm our shared wish for high levels of performance from all groups in American society. As the Supreme Court showed in the *Bakke* (Note 29) and *Weber* (Note 30)cases, it is quite possible to advocate a policy of equal results now and to implement it through the use of special preferences based on race, a strategy that can obviate the need to distort data in the service of policy preferences, if pursued consistently.

It is true that the adoption of that strategy by the Court in the absence of any legislative mandate to do so raises legal problems severe enough to convince many lawyers, myself included, that Mr. Justice Rehnquist's characterization of the *Weber* decision as Orwellian (Note 31) was fully justified. However, it is equally true that those decisions pose no problems at all for psychometricians as such or for the integrity and independence of the scientific tradition.

For scientists who would abandon that tradition in pursuit of their conception of social justice and equality, a cautionary tale from the recent history of science seems appropriate. The Russian biologist T. D. Lysenko and his allies were also pursuing an egalitarian ideal when they succeeded in giving their belief in the inheritance of acquired characteristics the force of law. The stubborn facts, however, remained unaltered and, as a result, they did not succeed in advancing the cause of equality by any definition. Instead, they succeeded only too well in retarding the progress of science by restricting the freedom of their fellow scientists in that country (Note 32).

If the new Standards Committee stands firm against pressures from within and from without, refusing to allow the 1981 Standards to depart from the tradition established in 1955 (Note 33)and continued through each succeeding edition, the battle lines between scientific standards and politicized agency guidelines will be clearly drawn. Of course, no one can be certain who will win that battle. However, the evidence available thus far suggests that a majority of the members of the current Court seem inclined, after a considerable amount of initial confusion, to respect the line between scientific findings and social policy choices (Note 34). For scientists to undercut the basis for such respect by blurring the line themselves would, I believe, be a sad development with unfortunate consequences, not only for psychometrics and for science as a whole, but for law and democracy as well.

In conclusion, I would like to thank you for allowing me to share these concerns with you. As most of you know, I am not a psychometrician but I have tried to follow developments in your field with some care, because I think all psychologists and all lawyers who respect the best in both traditions have a stake in the outcome of the current controversy over testing. Accordingly, I am pleased to have had the opportunity to address this distinguished group of psychometric experts.

Notes

1. Cronbach, L. J., and Meehl, P. *Construct Validity in Psychological Tests*. *Psychol. Bull.*, 1955, 52, 281.
2. Western thinkers, as Whitehead pointed out, pioneered in the development of science because they tended to assume the existence of order and stability in the universe. This assumption seemed natural to westerners because they were heirs to a theological tradition dominated by it, an outgrowth of the Judeo-Christian insistence that man is a rational creature, made in the image of a rational God. Eastern theology made no comparable assumptions and, as a result, eastern thinkers had no comparable faith that the universe was characterized by order and stability. *See* Whitehead, *Science and the Modern World*, 1925.
3. *See* James, *Some Problems of Philosophy*, 1911, 47-74.
4. *See* the letter William James wrote to his brother, Henry James, as he neared the end of his work on *Principles of Psychology*, in the collection of letters edited by his son: H. James (Ed.), *The Letters of William James*, 1920.
5. *See* Popper, *The Logic of Scientific Discovery*, 1968, 40.
6. *See, e.g.*, Judge Hand's remarks on what we do with juries in cases of negligence. "We say to them: 'What do you think is fair? What do you think is reasonable?' We call it a question of fact, but we have to close our eyes when we say it, for obviously it isn't." I. Dillard (Ed.), *The Spirit of Liberty: Papers and Addresses of Learned Hand*, 1960, 306.
7. *See* Cornford (Tr.), *The Republic of Plato* (10th ed.), 1954.
8. Galileo's controversy with the Inquisition is, perhaps, the best known, although as Whitehead justly comments, in the light of modern concepts of relative motion "Galileo's way of stating the facts was . . . in itself . . . not more true than the formulation of the Inquisition." Nonetheless, as Whitehead also notes, "Galileo's way of stating the facts was, beyond question, the fruitful procedure for the sake of scientific research." *See* Whitehead at 183, op. cit. at n.2 supra.
9. The power of the judiciary to invalidate Acts of Congress on the grounds that they violate Constitutional limits has no explicit basis in the Constitution itself. First exercised by the Supreme Court under Mr. Chief Justice Marshall in the case of Marbury v. Madison, 1 Cranch 137 (1803), that power is now firmly established in American law, although Thomas Jefferson rejected it as an undemocratic power and Mr. Justice Holmes mocked the claim that it was a necessary one.
10. Concern about the antidemocratic implications of this trend is widespread among leading Constitutional scholars. *See, e.g.*, Berger, *Government by Judiciary: The Transformation of the Fourteenth Amendment*, 1977, 314: "With Lincoln, I cling to faith in the ultimate good sense of the people: I cannot subscribe to the theory that America needs a savior, whether in the shape of a President or of nine—oft-times only five—Platonic guardians." *And see* Kurland, *Watergate and the Constitution*, 1978, 74: ". . . [I]n continuing to expand its own role in the government of the United States, the third branch again proved that we have arrived at the stage described by Raoul Berger in his new book *Government by Judiciary*. Sooner or later, this country must directly face the question whether it is prepared to entrust the judiciary with the mantles of Plato's guardians. The answer to that question will also determine the future of American democracy and, perhaps even of American liberty."

    See also Bickel, *The Supreme Court and the Idea of Progress*, 1978; Kurland, *Politics, The Constitution and the Warren Court*, 1970; and Lusky, *By What Right?* 1975. *And, compare* Cox, *The Warren Court*, 1968 with Cox, *The Role of the Supreme Court in American Government*, 1976.
11. *See, e.g.*, Bolce III and Gray, *Blacks, Whites and "Race Politics"*, *The Public Interest*, 1979, 54, 61; Gallup, G. H., *80% Majority Favors Ability as Criteria for Admissions*, Washington Post, November 20, 1977; Janowitz, *The Last Half-century: Societal Change and Politics in America*, 1978; National Assessment of Educational Progress, *Changes in Political Knowledge and Attitudes, 1969-76: Selected Results from the Second National Assessment of Citizenship and Social Studies*, 1978; Sheatsley, *White Attitudes Toward the Negro*, in Parsons, T. and Clark, K.B. *The Negro American*, 1967; Wilson, *The Declining Significance of Race*, 1978.
12. 42 U.S.C. §§2000 *et seq.*
13. *See* Mr. Justice Rehnquist's dissent in Kaiser Aluminum and Chemical Corp. v. Weber, 47 L.W. at 4861 for an excellent, detailed review of that legislative history. As he notes therein:

    > In most cases, '[l]egislative history . . . is more vague than the statute we are called upon to interpret'. . . Here, however, the legislative history of Title VII is as clear as the the language of §§703(a) and (d), and it irrefutably demonstrates that Congress meant precisely what it said in §§703(a) and (d)—that *no* racial discrimination in employment is permissible under Title VII, not even preferential treatment of minorities to correct racial imbalance.

    And, see McDonald v. Santa Fe Trail Co., 427 U.S. 273, 1976 for evidence that Mr. Justice Rehnquist's fellow justices initially shared his understanding that Title VII forbids discrimination against white citizens as well as against black ones. Note also that Mr. Justice Brennan's attempt to rationalize the glaring inconsistency between those two cases by stressing the fact that the employer's action in Weber was "voluntary" cannot withstand even the most cursory scrutiny: The employer's action in McDonald was equally so.
14. *See* Gorham, *Political, Ethical and Emotional Aspects of Federal Guidelines on Employee Selection Procedures*, in Schmidt and Hamner (Eds.), *Contemporary Problems in Personnel*, 1977, 159; *and see* the memorandum submitted to Deputy Attorney General Tyler by David L. Rose, Chief of the Employment Section of the Civil Rights Division of the Justice Department and published in the Bureau of National Affairs *Daily Labor Report* on June 22, 1976.

In explaining the need for the 1976 FEA Guidelines and the refusal of the EEOC to join with its sister agencies in endorsing them, Mr. Rose noted that:

> An unstated or covertly stated reason may underlie the apparent EEOC refusal to modify its present guidelines. Under the present EEOC guidelines, few employers are able to show the validity of any of their selection procedures, and the risk of their being held unlawful is high. Since not only tests, but all other procedures must be validated, the thrust of the present guidelines is to place almost all test users in a posture of noncompliance; to give great discretion to enforcement personnel to determine who should be prosecuted; and to set aside objective selection procedures in favor of numerical hiring.

See also the remarks of Eleanor Holmes Norton, Chairperson of the EEOC, at the Commission meeting of December 22, 1977:

> We do not see, however, comparable evidence that validated tests have in fact gotten black and brown bodies, or for that matter, females into places as a result of the validation of those tests. In other words, we do not see the kind of causal relation—that I think, when the great—and I regard it as a great new enforcement tool was discovered some years ago, we do not see quite the causal relationship we had expected to see. Rather we are faced with the possible anomaly that tests could be validated and have no effect or no appreciable effect as to minorities and women, and in particular minorities . . . So if the Commission, in effect, says to employers, as long as you validate your tests we're really not concerned about you anymore, I believe, in effect, it is saying that the presence of real people who are not in the work force is not as important as making sure that the tests have been validated. Therefore, I see some very positive advantages I must say in encouraging an employer to look at what the ultimate goal is. That is to say, did your work force have some minorties and females before the test was validated or does it have any appreciable number now that the test has been validated? And if you really don't want to go through that, but you are interested in getting excluded people into your work force, we would encourage you to do so.

15. *See, e.g.*, Anastasi, *Psychological Testing* (4th ed.), 1976; Bray and Moses, *Personnel Selection, Annual Rev. Psychol.*, 1972, *23*, 551, 553; Campbell, Crooks, Mahoney and Rock, *An Investigation of Sources of Bias in the Prediction of Job Performance: A Six-Year Study*, 1973; Cleary, *Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges, J. Educ. Meas.*, 1968, *5*, 115. Cleary, Humphreys, Kendrick and Wesman, *Educational Uses of Tests with Disadvantaged Students, Am. Psychol.*, 1975, *30*, 15-41, 88-96; Humphreys, *Adverse Impact on Blacks of Tests and Criterion Measures*, Paper presented at the American Psychological Association Convention, San Francisco, California, 1977; Humphreys, *Statistical Definitions of Test Validity for Minority Groups, J. Applied Psychol.*, 1973, *58*, 1; Hunter, Schmidt and Hunter, *Differential Validity of Employment Tests by Race: A Comprehensive Review and Analysis, Psychol. Bull.*, in press; Kendrick and Thomas, *Transition from School to College, Rev. Educ. Res.*, 1970, *40*, 151; Maier and Fuchs, *Effectiveness of Selection and Classification Testing*, U.S. Army Research Institute for the Behavioral and Social Sciences, Research Report 1179, September 1973; Maier, *Effects of General Ability, Education and Racial Group on Aptitude Test Performance*, Technical Research Note 228, Arlington, Virginia: U.S. Army Behavior and Systems Research Laboratory, 1971; Mitchell, *Predictive Validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for White and Negro Pupils, Educ. and Psychol. Meas.*, 1967, *27*, 1047; National Assessment of Educational Progress, *Functional Literacy, Basic Reading Performance*, 1976; O'Connor, Wexley, and Alexander, *Single Group Validity: Fact or Fallacy? J. Applied Psychol.*, 1975, *60*, 352; Schmidt, Berner and Hunter, *Racial Differences in Validity of Employment Tests: Reality or Illusion? J. Applied Psychol.*, 1973, *58*, 5; Stanley and Porter, *Correlation of Scholastic Aptitude Test Scores with College Grades for Negroes versus Whites, J. Educ. Meas.*, 1967, *4*, 199; Temp, *Validity of the SAT for Blacks and Whites, J. Educ. Meas.*, 1971, *8*, 245.

16. The Department of Justice, the Department of Labor, the Civil Service Commission and the Equal Employment Opportunity Commission.

17. *See* Uniform Guidelines on Employee Selection Procedures, 43 Federal Register No. 166, 38290-38315, §5B at 38298, §14D at 38303 and §16E and Y at 38307-8, August 25, 1978.

18. *See, e.g.*, Eyde, Primoff and Hardt, What Should the Content of Content Validity Be? Paper Presented at the APA Convention, New York, 1979.

19. *See, e.g.*, Lerner, *Democratic Values and Therapeutic Efficacy: A Construct Validity Study, J. Ab. Psychol.*, 1973, *82*, 491.

20. *See* Uniform Guidelines, op. cit. at n. 17 supra, §7B (3) at 38299 and §14B(8) at 38301-2.

21. American Psychological Association, American Educational Research Association and National Council on Measurement in Education, *Standards for Educational and Psychological Tests*, 1974.

22. *See* note 15 *supra*.

23. Id.

24. McKillip, Trattner, Corts and Wing, *The Professional and Administrative Career Examination: Research and Development*, Personnel Research and Development Center of the United States Civil Service Commission, Personnel Research Report, 1977, 77-1.

25. *See, e.g.*, Statement of Clifford I. Gould, Deputy Director, Federal Personnel and Compensation Division, to the Subcommittee on Civil Service of the Committee on Post Office and Civil Service of the U.S. House of Representatives, May 15, 1979.

26. Lerner, *Tests and Standards Today: Attacks, Counterattacks and Responses*, in R.T. Lennon (Ed.), *Impact of Social Change on Testing*, 1979.

27. *See, e.g.*, St. John, *School Desegregation: Outcomes for Children*, 1975.

28. Lerner, *Employment Discrimination: Adverse Impact, Validity and Equality*, in P. Kurland and G. Casper (Eds.), *1979 Supreme Court Review*, 1980; Lerner, *Equal Protection and External Screening: Davis, DeFunis and Bakke*, 1977 ETS Invitational Conference Proceedings, 1978; Lerner, *Washington v. Davis: Quantity, Quality and Equality in Employment Testing*, in P. Kurland (Ed.), *1976 Supreme Court Review*, 1977.

29. Regents of the University of California v. Bakke, 432 U.S. 265, 1978.

30. Kaiser Aluminum and Chemical Corp. v. Weber, 99 S. Ct. 2721, 1979.

31. 47 L. W. at 4858.

32. *See especially*, Z. Medvedev, *The Rise and Fall of T.D. Lysenko*.

33. American Psychological Association, *Technical Recommendations for Psychological Tests and Diagnostic Techniques, Psychol. Bull.*, 1954, *51*, 201.

34. Lerner, *The Supreme Court and the APA, AERA, NCME Test Standards: Past References and Future Possibilities, Am. Psychol.*, 1978, *33*, 915.

DISCUSSION

Norris: I got the impression you spoke of scientific decisions being made objectively, and you wanted to contrast that to the way, say, political decisions are made. I got the impression that by "objectively," you meant something like value free.

Lerner: I carefully didn't use words like "objectively" or "value free." I said that scientific decisions were made on the basis of empirical data, preferably a priori empirical data. I don't require personal neutrality. In fact, I find it hard to understand scientists who claim to have no personal preferences and no political views. I don't think that's possible. That's a whole other debate.

What I am talking about is the scientist's respect for data and for the scientific decision to base ultimate judgments on purely scientific questions on data alone, regardless of personal political preferences. That's why I used the Lysenko example. The best book I know describing the whole Lysenkoist period in Russian science is one by Zhores Medredev and I think most people here would agree with him very wholeheartedly in all of his statements and conclusions about science and about the damage done to it by its politization during Lysenko's reign. I do not think many people here would agree with him about political issues. He makes clear on every page of the book that he is a dedicated communist and a loyal Russian. He has no objectivity and claims none about social policy issues and choices. He is talking about science and about findings that all scientists must accept because of a shared respect for empirical data, not a kind of personal emptiness or whatever you want to call it.

Norris: I think that in making decisions about science, decisions about scientific matters, say, for example, which hypothesis to accept, scientists in those decisions use things that they value. They don't base the decisions solely upon the evidence. They base it upon the evidence plus other things like respect for simplicity, for example, which I think is a value. Some people think that, for example, the decision to throw away the old Greek view of the universe and to accept Copernicus' view was largely made upon simplicity considerations, not upon the data available because the data available didn't really favor one theory or the other. Both theories were adequate in explaining that data. Am I missing the point?

Lerner: I think so. If you talk about the data available at a particular point in time without recognizing that they may be adequate to resolve some questions but not others and without distinguishing between the two types of questions. That's why I tried to give two very different examples. My example of a question of the latter type was: shall we go with behaviorism as the guidelines tend to do, or take a less behavioristic or nonbehavioristic approach? Can we decide that today on the basis of the data? The answer is obviously not. It may take centuries to decide which approach is best.

In the meantime, can scientists take personal positions? Sure, and my own is a nonbehaviorist one. That's a value judgment. But that doesn't mean I won't also defend the right of my colleagues to make an equally legitimate interim scientific choice to be behaviorists. I will, because I think it is my obligation as a scientist to do that on questions of this type. On these questions, I think we can agree that the evidence is not adequate to compel a conclusion either way. And I think we can generally reach consensus on which issues cannot be resolved by data at the present time. I'm suggesting that this is one type of issue and that we mustn't let anyone outside the scientific community foreclose debate and research on issues of this type, issues that may take centuries to resolve.

There are, however, other types of issues that available data are adequate to resolve, at least for the present and the foreseeable future. I am suggesting that differential validity might be one of those issues, and that political considerations should not prevent scientists from resolving that issue and others like it, even though someday new evidence might emerge to alter our grasp of the problem.

Here, I am reminded that you spoke also about astronomical world views and in this realm I think Whitehead was right. He argued that by modern standards of relativity, Galileo was no more correct than the Inquisition that condemned him because, in fact, the answer here depends upon what stance you take in relation to the phenomenon; that is why both were right, in one sense. Still, I think there is no question—Whitehead thinks there is no question—that Galileo's approach was the fruitful one for science. His approach was a key factor in the development of Western science, and it would be deadly to let the law make decisions like that.

About parsimony, it may be that the Church's theory was better on that score and that many people, including scientists, might prefer a theory on those grounds, at least temporarily. But in the end, if the most parsimonious theory cannot explain relevant data that come up repeatedly, then scientists must recognize the data and discard the theory. Of course, as Clark Glymour pointed out, people are sometimes deceived by what they think the empirical data show, early on. But over a long period of time, when you have replication after replication and data and theory continue to conflict, then you can't pick theory over data. Of course you can; everyone here is free, up to a point. But you can't call yourself a scientist if you do that, because that violates the definition of science.

Norris: Scientists continue to use considerations like respect for the simple.

Lerner: In the interim, you are free to use anything you want to, as long as the data aren't compelling on one side or the other.

Norris: I am not saying the data aren't compelling, but you are compelled to make a scientific decision because there are other things than the data which go into making it. I am saying you are compelled to that decision, not by the data but by other things, like what the scientific community holds to be true at the time.

Lerner: I'm not sure I understand: do you think the field is not split between behaviorists and nonbehaviorists—that's a crude description but it will do for the moment—and do you think that you could not possibly get an agreement that both are scientifically legitimate at this point, even though we cannot now agree on which approach will ultimately prevail?

Norris: Not both are true.

Lerner: They are both not true? All right. We may all be wrong.

Norris: I didn't say they are both not true. I said not both are true and there is probably not enough information of any kind.

Lerner: Exactly my point. That's what I have been trying to say. But there is enough information, I think, to conclude that it is reasonable for us to defend one another's right to take either position until there are sufficient data to allow that choice to be made on an empirical basis. That's the point.

Schmidt: Is it your contention if behaviorists' assumptions were eradicated they would of necessity be replaced by something equally biased or at least biased in another direction? Is that your contention? Or are you arguing that guidelines on this issue can be neutral and should be; and, if so, how could they really be on an issue like this?

Lerner: I think it would be possible to write guidelines that made either approach acceptable. I don't think it's that hard to do, to at least approximate that ideal. As evidence of that, I would point to the professional test standards first issued by the APA in 1954 and to all the revisions of those standards up to the present. The standards are not perfect but I think they do a pretty good job, and have for a period of 25 years, in being fair to different sides. You know, the debate within psychology between behaviorists and nonbehaviorists hardly arose with the creation of equal employment opportunity legislation. It has been going on for a very long time and the standards, I think, have managed to be reasonably fair to both sides on that issue and on many other issues.

That being the case, I would question any need for separate government guidelines on test validity. Why duplicate the work of the standards committees? We don't have a separate set of government agency guidelines in other scientific fields or, at least, we didn't in the past. We are beginning to get more of them, more government intrusion into scientific affairs in violation of the tradition of independence for science in this country, and I think much of it is unnecessary and unwise. The government does not need to issue its own definitions of test validation in order to say that if your test has adverse impact, you must validate it. It can and should define test validation by simply telling people to look to the recognized scientific standards for that purpose. You don't need separate government definitions of scientific terms in most cases, unless the relevant scientific societies don't have any, have not done their own job, or unless you have some special, illegitimate political axe to grind that makes you want to change a scientific consensus or shift it askew in some way. The professional test standards, after all, don't give employers or anyone else a free ride. They aren't all that easy to follow. They are a tough set of standards, and they are not biased for or against any litigant. They weren't written to promote any particular political policy as far as I can tell. They reflect scientific consensus and are quite adequate for that.

Prien: Where would you place the Good Manufacturing Practices Guidelines? Certainly they aren't in psychology. How about the OSHA standards? Are those not comparable sets of criteria and standards for the evaluation of practice in other disciplines? It's a minor point.

Lerner: It interests me, too. I'm not sure how much it interests everyone else. I am not aware that industry had anything comparable to OSHA regulations, any industry-wide set of health and safety standards, so it's not as though the government were coming in to replace something that the groups involved had already done by themselves. It looks, at least to many people, like the government is filling a gap with OSHA. Whether they have done a good job or an extremely poor one is another question, and there are lots of dissatisfactions from many sources with those standards and with their costs in relation to their benefits. And there have been many battles lately which OSHA has been losing and there will probably be more, and there are major revisions underway. The point, I think, is that even when the government undertakes that sort of task with the best will in the world, it's a hell of a difficult job, and it usually gets into trouble. But, if private outfits won't, in effect, create reasonable standards for themselves, it tends to push the government into doing that.

I don't think that's the situation with psychometricians. I think they have behaved responsibly. They do have reasonable professional standards—not perfect, but pretty damn good ones—that reflect the general scientific consensus. So there isn't any hole the government has to fill. They are supplanting something, and that's a very different business. There are some interesting analogues, but I don't think the situations are really comparable. And of course, manufacturing is not a branch of science in the way that psychometrics is.

Wing: I think a more appropriate analogy might be the Guidelines on DNA research. That seems to be much closer to the kind of problem we are dealing with in terms of guidelines problems.

Lerner: I am not familiar enough with the DNA Guidelines to go into detail here but my impression is that there was much more participation by a much more representative sample of the scientific community in that. It was, I think, much more of a two-way job, a cooperative effort. The federal guidelines on employee selection were not. I mean, LABOR has how many psychologists working on this? Who is there besides Don Schwartz, and the EEOC had one, I think. Do they have two now? When I last checked, they had one. I don't mean to criticize those individuals but, you know, the APA, the AERA and the NCME never saw fit to have a committee of one or two. I think those organizations have chosen very distinguished psychometricians for their test standards committees for a long time. Nonetheless, they thought that a half dozen or more were needed to reflect a consensus. And I don't think that some of the federal guidelines on testing had full participation from even that many psychometricians, ignoring how they were selected and what instructions and burdens they worked under. So I think the DNA Guideline writers may have done a better job, but I'm not familiar enough with all of the relevant specifics to be sure.

McKillip: I don't think they were developed for the purpose of setting scientific standards. They were developed for the purpose of setting some kind of legal standard. And I would like to raise a small complaint about the APA, AERA, NCME Standards. I agree with you that they have been a pretty decent set of principles which scientists could live with and deal with. But, in being that way, they do become arguable enough, point by point, to cause or allow for considerable disagreement among scientists who have different scientific and social axes to grind. That's why you can find expert witnesses on both sides of a fairly clear-cut validation situation. Also, the standards are not standards; they are guidelines, and the guidelines are not guidelines; they are standards or regulations.

Lerner: I don't think the standards are perfect either. I voted with the Joint Review Committee to recommend that it was time for revision because I thought my colleagues on that Committee were right in concluding that revision was needed. But, as long as they are periodically revised and updated, I think they represent the best that can be done. The hope that we might write them in such a way that it would be impossible to debate them in an adversary system seems doomed to me. It's not possible to do or say anything that can't be debated in such a context. The question, as I understand it, is whether the standards are fairer and more reasonable than any set of politicized agency guidelines could be. And I believe that they are. I don't think the federal government is going to improve on them, certainly not by making them tougher when they are already too stringent for many practical purposes.

Of course I agree with you that the motives of most federal guideline writers aren't scientific. That's exactly why they shouldn't be writing them. When you say their motives are legal, however, I think that's questionable. The legal document that is controlling here is Title VII of the Civil Rights Act of 1964, section 2000(e) et sequelae. That is the law that governs and that is supposed to guide agencies like EEOC. That's what created EEOC and gave that agency its existence and its limited congressional mandate.

McKillip: I will withdraw that statement.

Lerner: I'm glad, because if you read the actual statute as Congress wrote it, it very clearly incorporates the traditional American legal attitude towards questions of scientific expertise. It doesn't invite EEOC or any of us to impose our social policy preferences on the nation, but it does express clear deference to the consensus of professional and scientific judgment in the field of testing. It doesn't do that in the kind of legal language which sometimes makes people from other professions despair. It does so in clear and plain English, and it does it repeatedly. EEOC, Labor and Justice didn't have a mandate, except to fill in a bit, perhaps, or clarify, but they weren't invited to redefine scientific terms and procedures or to make up their own more stringent and more biased ones. And the legality of that is very questionable.

Guion: One of the standards, I don't recall it by number, says that all criterion measures should be carefully examined for bias. This is a standard that was put into the 1974 Standards not by scientific process or consensus but by the social action attitudes of the various groups that the Standards were placed before in the process of writing them. It is a political statement more than it is a technical statement, primarily because we really don't know how to do it.

Now, if we were going to be limited in the next revision of the Standards to only those things for which there is a substantial data base, that particular standard would be deleted. Does that mean that it would be fair territory for government regulations to come in and dictate on how to determine whether a criterion measure is biased and to what degree?

Lerner: I think it's best for test standards to be uncorrupted in the sense that they reflect a real scientific consensus, not a pretend political consensus. If test standards are to be corrupted through politization, I think it's better it be done from outside and important for scientists to refuse to cooperate in doing that.

I have very strong social policy views myself, and I don't mean to discourage other scientists who do, whether they agree with me or disagree entirely, but I think such views belong in a separate document. I express mine by writing single-authored articles because I don't think there is any consensus within APA on social policy conflicts like the one between proponents of equal opportunity for all individuals forever and equal results for all groups now.

If I am wrong and it is possible to get a consensus on a social policy document in this area, then perhaps there should be such a document, clearly labeled as such and clearly distinguished from scientific documents issued by the same organization. I doubt that it is, myself, and I wouldn't want to be on such a committee because I think psychologists are too diverse a group to get consensus there, but I would encourage people who believe otherwise to try to create such a document and to try to get consensus on it. That would, I think, be better than trying to put political statements into the test standards themselves. I can't argue with the principal author of the 1974 Standards if he says the reasons for putting in one particular statement were partly political, but it seems to me that the scientific status of differential validity was at least debatable in the early 1970's. I think it's often, as I tried to say earlier, a question of timing. Something may look persuasive early and become less so later, as more and better data accumulate.

Thorndike: Isn't the standard one-sided? Don't you want your criterion variables to be so far as possible uncontaminated?

Novick: Isn't there scientific evidence to indicate that there is often bias in ratings and that it ought to be eliminated before those ratings can be used in test validation?

Lerner: The question of bias is, I think, a very broad one. There are many forms and types, many very different problems subsumed under that topic. I think the evidence against differential validity is now compelling and that evidence of differential impact is not per se sufficient to create an inference of differential validity. When the word bias is used more broadly and without definition in a general statement about criterion measures, I'm not really sure enough of what it means to be able to agree or disagree.

Guion: I'm not either. I do not think that it meant what Bob Thorndike was implying. I don't think it meant simply a freedom from contamination, although it should have. I think it was there with a special reference to racial and sexual bias, and I do not really know, or didn't, I think maybe I do now, but I'm not positive, but I certainly didn't then have any real idea how you tell the difference between a main effect and an interaction effect when you see the difference in the means of two groups on a set of criterion ratings.

Anastasi: This may be a simplistic reaction but it seems to me in that case, if you want to keep the standards scientific and apolitical, the thing to do is to define the bias. Certainly, we can't drop the word "bias." That's used all the time in sampling. But we should make more explicit what we mean by that in the sense of uncontaminated.

Lerner: I think that would be extremely helpful. Perhaps I confused people earlier. I wasn't suggesting that we abandon the word bias or the investigation of relevant forms of it. What I was referring to as unjustified was the requirement that people continue to collect data looking for differential test validities between races or ethnic groups on each test. I think people should be free to do that but I don't think we have a scientific basis for requiring it of everybody. I think there is clear evidence suggesting it is unlikely that there are substantial differential validities by group and that we shouldn't force everybody to continue searching for that because such a requirement can no longer be justified in scientific terms. Some people may want to continue the search for other reasons—fine—but we can't say to those who don't, you're unscientific if you didn't do that and we totally reject your data. I think that, by now, is unfair.

Novick: I find myself agreeing with almost all of what you have said, and any lack of agreement is on some minor details which we are not discussing. However, there are a couple of points that we might want to think about, apparently as a result of what you said earlier, which has got me a little more worried than I was before you started. We do have consensus, we had consensus on the Standards Review Committee as to what the document was going to be. The new Standards will not be political. It will be a scientific document and that was unanimous on the part of the Committee. I also agree with you completely that, to the extent possible, I would like to see the new standards preempt the guidelines. There may be a better word.

Lerner: At least for all scientists, and hopefully, for the Court as well.

Novick: There are two points, I think, to be made. One, I think it's well recognized that professional organizations, whether it be APA, AMA or any others, are generally ineffective in enforcing their own standards. I am not quite convinced of this and I guess I've changed my view as to the possibility of APA or AERA or NCME going much further than they have gone to this point in enforcement which, of course, then leaves open the question of enforcement, and in that sense, probably leaves it with EEOC or perhaps some other agency, as politics change.

The thing that has concerned me, the new issue that arises, is your statement towards the beginning of your presentation that the question of proof is very different in the legal system than it is in the scientific business, and yet you and I both are looking for this scientific document to be the basis for discussions of proof in a generally legal adversarial system. Now, this is what's concerning me and maybe you can help me out.

Lerner: I would like to try, but I'm concerned that we're running overtime here. I have spelled out my views on that problem in some detail in an article in the 1979 Supreme Court Review, in press, and I'm afraid the best I can do is give you a very inadequate, brief summary of that fairly elaborate argument here.

The main point is that legal standards in every other area are and should be less stringent than scientific standards, and that should be the case with regard to testing also. I want to see professional standards used when necessary, but I don't want to see them blindly enforced so that every small businessman or professional who hires a few people has to do a full scale local validation study. These people don't have the money or the staff to do full scale studies and they should have a chance in court to justify using a test which is arguably reasonable for their legitimate purposes without going through a full scale validity study in every instance. Because if we insist on formal validity in these situations, all we are saying is that small employers can't use tests at all. We are forcing them to use other alternatives which are even less valid. In other words, legal standards should be less stringent than the scientific ideal expressed in the APA, AERA, NCME document. They should never be more stringent, as the guidelines have been. What Title VII and all other such statutes require of people is reasonableness, not perfection. Compliance with the standards should be one way of demonstrating reasonableness, but there must be other, less stringent ways for test users to do that, and that is what I tried to spell out in the article I mentioned.

JOHN E. HUNTER

CONSTRUCT VALIDITY AND VALIDITY GENERALIZATION

I have come not to present theories but to verify them. Recent fact has supported not the pessimistic theories of the 1970's but the optimistic theories of the 1930's. The gloom and doom of the 1970's has been shown to be the reification of sampling error. The sophisticated methodologies of the 1970's are superstition.

How have the false beliefs of recent times arisen? The hundred years of psychometric research from 1850 to 1950 was a period of steady advance in knowledge and a steady increase in optimism concerning the usefulness of cognitive tests. Visual ability tests were found to predict the performance of tool and die workers, perceptual speed was found to predict the performance of clerks, etc. Most obvious hypotheses about the cognitive processes used in tests were verified by relating test scores to corresponding tasks in industry and in the military. Indeed, the successes of the first hundred years generated an explosion of research that has been accompanied by a steady growth of pessimism and the propagation of corresponding stringent methodological requirements. Most of these requirements have no basis in empirical or theoretical findings and are frequently contrary to existing knowledge. For example, many psychologists believe that correction for attenuation is a suspicious process of doctoring the data. Others feel that correction for restriction in range must be immoral since it increases the size of correlations, and is thus self-serving. A large number of psychologists accept each of these corrections separately but feel that to do both is to be subject to naive self-aggrandizement. Even the occasional empirical verification of these formulas has not stemmed the tide of disbelief stemming from the pessimism of the 1970's. But these are familiar issues. Let me turn to relatively unchallenged myths. For example, many psychologists believe that the correlation between test scores and job performance will be higher if the psychologist in charge has done a detailed job analysis. But a job analysis changes neither test scores nor job performance so the mechanism for such a recommendation is unknown. Nor is there a shred of data to support this belief. As a second example, consider recommendations regarding supervisor ratings. The same narrow behaviorism which has been discarded in experimental psychology has been picked from the garbage can and paraded as a panacea for the problem of rater halo and the alleged problem of rater bias toward ethnic groups. Yet all existing empirical evidence has been exactly contradictory to such claims. Behavioral anchored rating scales whether specific or general in nature do NOT reduce halo. Halo can only be reduced if the job can be set up so that more than one rater becomes familiar with each worker.

The psychological study of stereotypes and prejudice has often found that such beliefs are supported by a "kernel of truth," some fact which is misinterpreted as a basis for the belief. For example, crime rates are higher in black neighborhoods in Chicago than in white neighborhoods. The bigot falsely derives from this fact the inference that all blacks are untrustworthy. Is there a similar "kernel of truth" for the false belief that tests are frequently invalid predictors of job performance? The kernel of truth in this case is the typical review study published over the last 20 years. The review study says that some people have found the test to be valid and other people have found the test to be invalid. The study then endorses the "theory" of situation-specific validity by concluding that there are some unknown variables which cause the test to be valid in some settings but not valid in others. That is, there must be some unknown moderator variable which interacts with test and job performance in some way. Thus, even though the study may cite 200 or more studies on a given issue, the author will conclude that further research is necessary!

I will show below that the primary cause of poor review procedure is the statistical significance test. The significance test has lulled us into the belief that it is possible to eliminate the sampling error in single studies and has thus led reviewers to consider each study on a one-by-one basis. The appropriate answer to sampling error is cumulation of descriptive statistics across studies as recommended by Glass and his associates and by Schmidt, Hunter, and their associates (work reviewed in Hunter, Note 2). The methods derived for validity generalization (Hunter, Note 2; Schmidt, Hunter, Pearlman, and Shane, 1979) can be used to assess the extent to which variation from one study to the next is due to sampling error or due to real differences in results. The appropriate solution to the problem of sampling error in single studies is not the significance test but the confidence interval. Most psychologists have greatly overestimated the amount of hard information there is in small-sample studies.

This paper will do three things: First, it will review the literature on cumulative findings in test validation in employment research. The bottom line here is that the myth of test invalidity is entirely an artifact of sampling error. Second, a mathematical argument will be presented showing that the statistical significance test produces invalid reviews of cumulative research and hence should be abandoned in favor of the confidence interval. Third, these findings will be related to the general debate over the usefulness of constructs such as cognitive abilities. Behaviorist arguments will be refuted both empirically and philosophically.

Validity Generalization

How small must the sample size be in order to be "small"? That depends on the size of the desired confidence interval. For a population correlation of .40, the sample size required to estimate the correlation to within ±.05 is 1085. If the sample size was as low as 400, the width of the confidence interval would be ±.08. If the sample size was as low as the average reported for test validation studies by Lent, Aurbach, and Levin (1971), i.e. as low as 68, then the width of the confidence interval would be ±.20. Thus, any sample of less than 400 or so would be "small" for most purposes and sample sizes of 1000 or more would be required for the detection of subtle differences. Only cumulation of results across studies can generate such sample sizes in most areas of contemporary research.

The grim facts above were pointed out by Schmidt, Hunter, and Urry (1976) in the area of test validation studies. That this is no purely theoretical problem can be seen from Table 2 of the Lent, Aurbach, and Levin (1971) study of 1506 correlations in the test validation literature. In that table there is a correlation of .77 between sample size and the proportion of correlations found to be statistically significant. Indeed, this result suggests that the vast majority of findings of "invalid tests" might actually be an artifact of small samples.

How much of the variability in results across studies is due to artifacts such as the use of small samples? We have sought to answer exactly this question in the field of test validation studies (Schmidt and Hunter, 1977; Schmidt, Hunter, Pearlman, and Shane, 1979; Pearlman, Schmidt, and Hunter in press; Schmidt, Hunter, and Caplan, Note 3; Schmidt, Rosenberg, and Hunter, in press). In the typical test-job combination, we find that more than 50 percent of the variation across studies is accounted for by three artifacts: sampling error, variation in criterion reliability, and variation in restriction of range. Moreover, it is the sampling error implicit in the use of small-sample studies which accounts for 85 percent of the artifactual variance. From the frequency of outliers in our distributions, we have become convinced that the remaining variability left in our distributions after we take out artifacts is so small as to have no practical or theoretical significance within the field of test validation.

A Proposed Moratorium on Statistical Significance Tests

Consider the typical review study in the social sciences. Many studies are found relating variables X and Y, with a long list of authors finding a significant relation between them and a long list of authors finding no significant relation between them. Thus, there are "conflicting results" which can only be reconciled with "further research designed to tease out the subtle interactions and methodological deficiencies which must explain such discrepancies." Our validity generalization studies show that such review studies can be just flat wrong. The unspecified "interactions" and unproven "methodological deficiencies" may be ghosts summoned by the use of small samples and the reification of sampling error.

Consider an example. According to 144 studies, Pearlman, Schmidt, and Hunter (in press) found the correlation of general intelligence and proficiency in clerical work to be .51. That is, if a perfect validity study were done using the entire applicant population, using a perfectly measured job proficiency measure, then the correlation between intelligence and performance would be .51. But proficiency measures cannot be obtained on applicants. Performance can be measured only on those who are hired. Most organizations hire fewer than half of those who apply. Suppose that those hired are those in the top half of the distribution on intelligence. Then because of restriction in range, the correlation between test and performance will only be .33 rather than .51. But it is also impossible to obtain perfect measures of job performance. Typically, the best feasible measure is the rating of the single supervisor who knows the person's work well enough to rate it. According to the review of King, Hunter, and Schmidt (in press), the correct interrater reliability of a single rating by a single supervisor would be only .31 (though the coefficient alpha of judgments by that judge would be a spurious .62).

If the rater were asked to make multiple judgments and those ratings were combined into a composite score, then the correct interrater reliability would rise to .62 (though coefficient alpha on ratings by a single judge would misleadingly read 1.00). In this "best" of cases, the reliability of the job performance measure would be .62 (actually less, though we will ignore the reduction in reliability produced by restriction in range) and the potential correlation between test and performance would be dropped from .33 to .26.

So because of problems inherent in doing field studies, the investigator begins work with an underlying population correlation of .26. What are the implications of doing a study with a small sample? Suppose that Smith does a study with the minimum sample size recommended by the EEOC, i.e., $N = 30$. Then if the expected correlation of .26 were found, it would not be statistically significant and Smith would label the test as "invalid." But suppose that Jones found the same correlation of .26 using the mean sample size reported by Lent, Aurbach, and Levin (1971), i. e., $N = 68$. Then the same correlation would be statistically significant and Jones would label the test as "valid." Here we have a prime example of false "conflicting results in the literature." The same correlation can be labeled "significant" in one study but "nonsignificant" in another. Most reviews treat all studies alike and completely ignore the fact that studies with different sample sizes have a completely different meaning for "significant."

But suppose that all studies were done with the same sample size. Would reviews find the research literature consonant with the uniformity which we have defined our example to have? Certainly not. We have assumed that the observed correlations would equal the population correlations, but this is quite false. If all studies were done with 50 subjects, what would the effect of sampling error be? For 50 subjects, the correlation must be .28 to be significant at the .05 level using a two-tailed test. Given a population correlation of .26, the observed correlation will have a mean of .26 and a standard deviation of .13. The probability that the observed correlation will be larger than .28 and hence be labeled as "significant" is .44. Thus across studies, the test would be correctly labeled "valid" 44 percent of the time and would be falsely labeled "invalid" 56 percent of the time. The author list in the review on one side of the issue would be just about as long as the author list on the other side of the issue. Note that all authors who claim the test is "invalid" are actually wrong. That is, the conclusion stated in the majority of studies is wrong. Thus the "preponderance of data" using counts of statistical significance can be completely false.

If the statistical significance test yields incorrect results when applied across studies, then what is a reasonable alternative? The long range solution is cumulation across studies. The short range solution is the confidence interval. Suppose that Jones obtained an observed correlation of .15 which is not significant while Smith obtained a correlation of .37 which is significant. Using significance tests, they would conclude that there was some unknown difference in the jobs in their two studies which caused the test to be valid for Smith but not for Jones, a typical false conclusion. But suppose that each had used confidence intervals. The confidence interval for Jones is −.13 to +.43 (i.e., .15 ±.28 for N = 50) while it is +.09 to +.65 (i.e., .37 ±.28) for Smith. These confidence intervals have an overlap of .09 to .43 which includes the population value of .26. They would certainly perceive their results as being comparable rather than as conflicting.

Most psychologists believe that statistical significance tests are a solution to the problem of sampling error. Nothing could be farther from the truth. Statisticians have long known that there is NO solution to the problem of sampling error within single studies. Contemporary use of significance tests is a travesty of statistical theory, since Type II error—the most frequent type—is completely ignored. Statistically sophisticated psychologists have tried for at least 25 years to state this message in terms of the "power" of statistical tests, but I have never seen a shred of evidence in the review studies that this message has any meaning to substantively oriented researchers. The time has come to dump the statistical test in favor of what has always been the more powerful and more meaningful alternative: the confidence interval.

Dust-bowl Empiricism

The attack on constructs has come primarily from two groups: the dust-bowl empiricists and the behaviorists. The dust-bowl empiricists believe that the only important outcome of any endeavor is an increase in the extent of prediction of some important fact in applied psychology, especially job performance or academic achievement. For them the question is: what good are all these theories and constructs if they don't lead to the prediction of important behavior? If our theories about abilities cannot lead to correct predictions about what tests will predict performance in given jobs, then of what use is the concept of ability? For the dust-bowl empiricist, the apparent variation in test validity across jobs suggested that current theory was useless and should be abandoned. The work on validity generalization has now shown this belief to be based on sampling error.

But the work on validity generalization is still limited to a relatively small number of jobs and does not yet reply to the more general criticism: Since thought and problem solving enter into practically all work, then cognitive ability tests should be valid for nearly all jobs. The same data which led to the doctrine of situational specific validity also seemed to suggest that there are many situations in which cognitive ability tests are not valid. But there is important empirical evidence on this point gathered by Helm, Gibson, and Brogden (Note 1). This evidence is presented in Table 1.

The data reported in Table 1 are based on 70 independent validity studies each done on sample sizes that average 350 persons. The studies were organized to provide exact cross validation on each of 35 jobs. Thus averaging across the two studies for each job yields a correlation based on 700 subjects. Each row of Table 1 represents an independent sample. This unique study provides a broad study of the validity of the most common cognitive ability tests which is virtually without sampling error.

Are cognitive ability tests generally valid as would be predicted from the theories of the 1930's? For nearly all the jobs listed, there is an ability test which has a validity of .60 or more. The least predictable job is high-speed radio operator, a relatively specialized psychomotor task. Yet the validity of the common ability tests is high enough to be very useful: .22, .19, .16, .26, .21 for vocabulary (i.e. verbal ability), arithmetic reasoning, spatial aptitude, mechanical aptitude, and clerical speed respectively; es-

TABLE 1
Validity Coefficients of ACB Tests
for 35 Army School Courses
Group A Samples

			Vocabulary	Arithmetic Reasoning	Spatial Aptitude	Mechanical Aptitude	Clerical Speed	Radio code Aptitude	Shop Mechanics	Automotive Information	Electronics Information	Radio Information
			ACB Tests									
Current MOS	Course Title	N	RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
271-2	Fixed Sta Radio Rep	210	58	69	58	61	38	52	57	42	67	62
281	Microwave-Mult Channel Rep	216	36	52	41	48	23	16	44	40	55	54
282	Radar Rep	242	38	46	40	55	22	24	32	38	50	48
296	Field Radio Rep	280	54	62	60	57	39	42	59	50	63	48
320	Field Wireman	330	45	49	48	43	31	32	52	43	58	40
403	FC Instrument Rep	214	52	49	58	59	34	31	54	50	52	37
411	Ammunition Supply	767	54	55	45	42	44	28	50	42	38	20
421	Arty Mech-Light Weapons	196	64	64	57	58	54	29	63	60	49	13
422	Small Arms Weapons Mech	418	53	60	60	61	48	34	69	64	44	26
424	Turret Arty Rep	183	66	70	60	62	52	41	66	61	44	28
442	Welder	235	55	50	59	59	37	33	56	56	38	19
443	Machinist	296	38	59	57	58	33	30	66	46	46	34
452	Dental Lab	100	38	44	52	60	24	29	53	40	40	28
631	Automotive Mech	154	31	48	45	57	28	17	51	64	45	28
632	Track Veh Rep	448	62	64	56	67	47	34	69	69	49	27
632	Armor Track Veh Maint	248	48	56	54	65	36	36	58	63	56	34
634	Fuel & Elec Systems Rep	523	51	61	56	70	22	35	63	71	61	45
635	Track Veh Chassis Rebuild	103	33	48	52	60	42	47	64	62	56	38
710	Clerk	311	62	68	39	43	52	42	41	20	44	30
712	Stenography	569	49	50	38	33	46	32	24	16	28	15
714	Postal Operations	293	65	71	51	48	52	39	47	36	38	26
216	Personnel Admin	286	59	66	43	45	55	40	52	26	45	21
716	Personnel Mgt (Enl)	556	67	70	48	49	53	39	44	20	40	30
717	Adv Army Admin	406	65	63	44	46	53	26	35	34	37	31
753	Machine Acctg	383	60	69	59	50	53	45	39	27	47	33
763	Ord Storage Spec	291	69	75	50	58	57	36	65	45	44	33
911	Medical Aidman, Adv	303	61	61	57	52	44	50	48	34	50	41
912	Medical Tech	271	37	26	17	28	20	16	19	21	26	13
917	Dental Asst	367	51	54	44	41	29	27	38	22	49	37
941	Cook	305	37	39	34	34	32	28	36	33	28	12
951	Military Police, Enl Adv	159	61	67	51	52	39	32	49	24	44	38
952	Disciplinary Guard, Enl	144	52	59	41	52	24	32	48	47	45	33
953	Criminal Investigation	192	62	63	51	57	42	24	41	31	54	49
051	**Radio Op (Interm Speed)**	150	22	18	12	19	22	32	07	13	16	09
052	Radio Op (High Speed)	233	23	27	12	31	23	48	16	30	27	22

pecially if these are combined in a multiple regression equation. Moreover, it was possible for psychologists to create a specialized radiocode aptitude test which correlates .42. Indeed, cognitive ability tests are generally valid as predicted.

Furthermore, the variation in validity is as would be expected. The high validity coefficients tend to be for jobs with a high degree of problem solving such as repairmen, while the lower validity coefficients tend to be found for those who carry out fixed procedures such as cook or dental lab technician. Specific information tests have more general validity than might be naively imagined, though the wide validity of these tests probably stems from the correlation between specific information and general abilities.

Philosophical Behaviorism

Most current psychologists believe that personnel selection will not work in a given context unless it is based on a behavioristically organized analysis of the fine details of job performance. There is no evidence to suggest that these detailed job analyses yield any improvement in se-

TABLE 1 (cont.)
Validity Coefficients of ACB Tests
for 35 Army School Course
Group B Samples

			Vocabulary	Arithmetic Reasoning	Spatial Aptitude	Mechanical Aptitude	Clerical Speed	Radio code Aptitude	Shop Mechanics	Automotive Information	Electronics Information	Radio Information
							ACB Tests					
Current MOS	Course Title	N	RV	AR	PA	MA	ACS	ARC	SM	AI	EI	RI
271-2	Fixed Sta Radio Rep	323	50	61	55	61	37	47	54	44	71	70
281	Microwave-Mult Channel Rep	236	37	41	44	36	16	21	35	18	58	56
282	Radar Rep	237	34	49	49	47	21	37	47	28	63	56
296	Field Radio Rep	481	41	51	38	50	39	46	40	55	55	
320	Field Wireman	330	59	66	54	51	53	42	60	42	56	50
403	FC Instrument Rep	214	36	54	55	65	26	35	66	56	56	38
411	Ammunition Supply	702	53	51	43	49	45	25	47	37	27	12
421	Arty Mech-Light Weapons	205	64	65	66	67	60	34	67	64	46	19
422	Small Arms Weapons Mech	430	62	67	62	60	50	38	65	62	46	27
424	Turret Arty Rep	186	61	63	52	59	62	39	61	56	46	24
442	Welder	243	38	52	44	48	32	27	50	45	34	25
443	Machinist	314	50	61	62	55	44	28	62	52	52	40
452	Dental Lab	121	11	26	36	33	-03	18	27	31	17	25
631	Automotive Mech	198	57	61	52	66	36	22	60	72	57	49
632	Track Veh Rep	430	65	65	56	67	47	23	72	78	53	16
632	Armor Track Veh Maint	248	45	63	57	59	35	37	60	63	47	24
634	Fuel & Elec Systems Rep	522	55	64	62	69	35	35	69	70	60	35
635	Track Veh Chassis Rebuild	114	50	50	46	56	37	12	51	50	47	30
710	Clerk	340	67	75	56	54	59	43	48	34	52	34
712	Stenography	569	49	51	39	26	49	36	18	03	23	22
714	Postal Operations	295	61	67	53	44	49	40	41	29	38	26
716	Personnel Admin	286	62	65	46	42	45	35	38	21	39	27
716	Personnel Mgt. (Enl)	556	70	69	50	54	54	40	45	19	54	29
717	Adv Army Admin	401	63	67	45	47	53	34	41	22	47	34
753	Machine Acctg	355	60	60	50	49	49	38	45	32	50	33
763	Ord Storage Spec	274	70	75	66	67	60	44	63	50	49	18
911	Medical Aidman, Adv	200	67	67	41	45	45	32	46	17	47	40
912	Medical Tech	200	27	30	25	31	27	26	28	23	21	18
917	Dental Asst	200	65	65	49	48	43	43	36	19	49	41
941	Cook	338	28	32	25	26	28	28	30	26	21	06
951	Military Police, Enl Adv	363	60	63	55	51	47	34	51	32	35	25
952	Disciplinary Guard, Enl	139	66	67	44	54	57	39	49	36	43	20
953	Criminal Investigation	106	75	79	67	58	53	60	38	-05	53	51
051	Radio Op (Interm Speed)	145	43	37	34	39	39	59	37	35	16	14
052	Radio Op (High Speed)	233	21	11	20	22	19	36	17	11	17	27

Note—Decimal points omitted in table. These data are from Helm et al. (Note 1)

lection over the crude categorization of jobs into classes such as "clerical", "drafting," etc. Indeed, there is growing evidence that the validity of a given test is invariant across broad job categories. However, behaviorism has never been an empirical doctrine but rather a philosophical position and is best attacked on such terms.

The philosophical doctrine of behaviorism was strongly felt in psychology for 50 years, though its influence was slow to permeate industrial and educational psychology. However when the doctrine of the situational specificity of validity seemed to have destroyed the empirical basis for faith in the existence of traits, skills, and abilities, behavioristic ideas were quick to fill the vacuum. Behavioristically oriented psychologists argued that the trouble with the notion of "ability" is that it depends on the notion of "internal processes" and that such "unobserved" and "unobservable" events can never be the basis of an empirical science. Instead they proposed basing personnel selection on observed "behavior;" a fine-grained analysis of the details of job performance.

There are two central claims to the modern behavioristic beliefs in the field of job analysis: that "abilities" are not observable and that "behavior" is observable. Both claims are false to fact. Consider the supposed observability of behavior. For example, suppose that a worker is to screw a certain bolt into a certain hole in each automobile as it passes him on the assembly line. Is "screwing in the bolt" an observed behavior in the worker? Certainly not. During the nineteenth century, physiologists and psychologists established the fact that no event can have an effect on the brain unless it is mediated by afferent nerve fibers connecting some sensory mechanism to the brain. Thus the bolt hole in the automobile cannot be the causal determinant of behavior per se; it must be a pattern of afferent nerve impulses in the brain (i.e. an unobserved event). Moreover, a detailed analysis of the neural pathways that are implied in vision alone shows that no pattern of neural impulses is *ever* repeated. There is such fine detail of neural pathways that the brain records every least difference in the angle of the head, the viewing distance, the position of the car on the assembly line, the difference in the postural muscles stemming from differences in foot position or tool position, etc. Thus, strictly speaking, the stimulus of "bolt hole" is an event in the observing psychologist's head, not in the eye of the worker. If there is to be an equivalence of successive bolt holes for the worker (and we have no doubt that there is), then that equivalence is a perceptual *construction* of the meaning of similar events computed in the brain. Thus the "stimulus" of behaviorism is a hypothetical construct whose existence is a matter of inference from observed data based on an (implicit in this case) theory. Thus, "stimuli" are no more "observable" than are traits.

Furthermore, physiology has established similar facts about motor patterns. A motor act such as "screwing in a bolt" is known to be a highly complex pattern of time-sequenced patterns of neural impulses to thousands of muscle fibers, to postural muscles, to eye muscles, etc. Thus, it is well known that no "response" ever repeats itself, either. Thus, any equivalence of successive acts must be an internal perceptual process carried out by the brain. Thus, strictly speaking, "responses," too, are events in the observing psychologist's brain; and hence not directly observable. The belief that workers react to successive acts as "repetitions" shows that "behavior" is a hypothetical construct which is assumed stable in current theory and which has been verified using observations interpreted within that theory. Thus, the "response" of behaviorism is no more an "observable" than is a skill; it, too, is a hypothetical construct in the minds of those who use it as a theoretical device.

Is "quantitative ability" an unobserved variable whose existence depends on some nebulous theory of unobserved brain function? Certainly not. So far as we know, not one test being used in current industrial selection was derived from a psychological theory of numerical or quantitative ability (such as those based on Piaget's theory of operations which are under study in educational psychology). Rather the basis for existing quantitative ability tests rests entirely on what is currently called "content validity" (see Ebel, 1977 for an extensive discussion of this fact for all the ability tests); i.e., quantitative ability is *defined* in terms of the ability to correctly carry out a certain family of tasks. Although this family of tasks has never been written out in full formal manner (to our knowledge), it is sufficiently well specified that many different investigators in different places, using different populations, with different applications in mind, and at different points in time have constructed instruments which have subsequently proved to be statistically equivalent to one another (i.e., to define the same general factor to within error of measurement). In hundreds of cases, jobs such as accountant or computer programmer have been predicted to depend on quantitative ability because quantitative processes are obviously part of the job, and in hundreds of cases these predictions have been borne out. So "quantitative ability" is an empirically defined measurement procedure like length or weight. Though no measuring instrument in science is ever perfect (indeed our measurement theory of quantitative ability says only an infinitely long test can be perfect), the values can certainly be said to be "observable" in the usual language of science.

In summary . . . strictly speaking, the word "ability" like the word "behavior" represents a hypothetical construct whose nature is defined by a psychological theory. Observed events and values can be correlated to these words only to the extent that the corresponding theoretical sentences have been shown to be valid. On the other hand, the theory that certain acts are repetitions of one another and the theory that individual differences stem from psychological skills or abilities are theories which predate psychology and which have mountains of empirical support. Thus in ordinary language, both can be regarded as directly "observable."

General Conclusion

Once personnel psychology seemed to be a basically simple affair. To predict performance in a job required two steps: (1) identify the abilities which determine individual differences on that job and (2) construct tests to measure those abilities. The final selection instrument would then be a composite test (or composite score) based on the identified abilities. It was thought that both steps could be carried out by psychologists as a group for job categories such as "clerical work." It was believed that the day was soon coming when a psychologist could simply look to the literature to tell an organization what test to use to select for any given job. Over the last twenty years, this belief has come under severe attack from two directions: First, the data on selection seemed to support a doctrine of job-situation specificity which was finally crystallized by Ghiselli who asserted that the factor structure of jobs varies so much from job situation to job situation that every employment context demands a new validity study. Second, the philosophy of behaviorism caused many investigators to feel uncomfortable about "hypothetical constructs," such as ability, and to feel that selection must be grounded in a detailed analysis of the behaviors exhibited on the job.

Recent work has come full circle to original belief about the relationship between ability and job performance; the elimination of methodological artifacts in past work is revealing an underlying structure of exceeding simplicity. The basis already exists for devising a data collection scheme to permit the empirical validation of different abilities for different job families. In fact there has already been enough data gathered to complete most of that structure, though data reporting has been so incomplete that only a small portion of that work can be salvaged for this purpose.

Notes

1. Helm, W.E.; Gibson, W.A.; and Brogden, H.E. An empirical test of shrinkage problems in personnel classification research. Adjutant General's Office: Personnel Research Board Technical Report. Note 84, October 1957.
2. Hunter, J.E. Cumulating results across studies: a critique of factor analysis, canonical correlation, MANOVA, and statistical significance testing. Invited address, American Psychological Association, New York, 1979.
3. Schmidt, F.L.; Hunter, J.E.; and Caplan, J.R. Test of a new model of validity generalization: results for two occupations in the petroleum industry. Washington, D.C.: The American Petroleum Institute, 1979.

References

Bartlett, C.J.; Bobko, P.; Mosier, S.B. and Hannan, R. Testing for fairness with a moderated multiple regression strategy: an alternative to differential analysis. *Personnel Psychology*, 1978, *31*, 233-24?.

Ebel, R. Comments on some problems of employment testing. *Personnel Psychology*, 1977, *30*, 55-63.

King, L.M.; Hunter, J.E.; and Schmidt, F.L. Halo in a Multidimensional forced choice performance evaluation scale. *Journal of Applied Psychology*, in press.

Lent, R.H.; Aurbach, H.A. and Levin, L.S. Research design and validity assessment. *Personnel Psychology*, 1971, *24*, 247-274.

Pearlman, K; Schmidt, F.L.; and Hunter, J.E. Validity generalization: results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, in press.

Schmidt, F.L. and Hunter, J.E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 1977, *62*, 529-540.

Schmidt, F.L.; Hunter, J.E.; Pearlman, K. and Shane, G.S. Further tests of the Schmidt-Hunter bayesian validity generalization procedure. *Personnel Psychology*, 1979, *32*, 257-281.

Schmidt, F.L.; Rosenberg, I.G.; and Hunter, J.E. Application of the Schmidt-Hunter validity generalization model to computer programmers. *Journal of Applied Psychology*, in press.

Schmidt, F.L.; Hunter, J.E., and Urry, V.E. Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, 1977, *62*, 529-540.

DISCUSSION

Novick: We have a proposition A, that there is enormous situational variability in prediction, both validity and perhaps regression. We have a proposition B, that there is little or no variability. And the paper today takes us from A all the way over to B.

I think that I am convinced by this paper—but perhaps was convinced before this—that A is surely not correct. I am yet to be convinced that B is completely correct. I think there are certain kinds of situations or certain situational specificity, certain areas of aptitude treatment interaction where there are parallels, I think.

The difficulty here in deciding just how far toward B one gets is, I think, to some extent a methodological problem. I certainly couldn't agree more that we ought to avoid hypothesis tests, and certainly confidence intervals are much to be desired, though one might think of some better things to do.

Where I think we run into a problem with what we have heard today—two problems, in fact—one is focusing attention on correlation coefficients. There is a society—not a real society—but a society formed for the abolition of correlation coefficients, and I think I subscribe to the position of that society. I think we ought to be looking at regression lines to answer these questions and there I think we will find some differences in regressions sometimes.

The other difficulty with the approach taken here is that somehow we think that we can use multiple corrections for this, that, and the other thing, and expect to come out with something close to the truth. I am generally suspicious of the use of even one restriction, restriction of the range, when there is much of a correction to be made. When we correct two or three things at one time, I lose faith.

Nevertheless, I think you have taken us down the right road—B—perhaps a bit too far. And I think perhaps we need a slightly different vehicle to find out just where we all stop.

Hunter: Let me respond to as many of those points as I can remember. First of all, I have not taken you down that road in this paper. This paper does not summarize the data, and it was neither I nor my opinions that brought us down this road. I think we are up to about four or five successive papers where we have been looking at massive data we accumulated in the various areas, and it is the data that give the results and not my philosophical predilections.

The next thing I would like to say is that it is not necessary that there be no variation in validity from study to study. What is required is that the variation be too small to be of significance for practical decisions. Now, the practical decision to be made is, is the test valid or not valid? Will the use of the test bring economic benefit and other kinds of benefits to the firm that uses it?

For that, all we need to know is that the probability that the test would be invalid is vanishingly small, and that is certainly true. In other words, the degree of variability, once you remove certain things, is much, much too small to have practical significance.

Theoretically it may be important that validity be .40 in one setting and .50 in another setting after correcting for restriction in range and unreliability. Such a difference might mean a significant difference in management practices, for example. However, from a practical point of view and the point of view of the *Uniform Guidelines*, in terms of setting up an instrument which is supposed to dictate reasonable procedures for employers to be using, I think the evidence is clear. There is no substantial degree of variability in validity across jobs from setting to setting, or across times. So, therefore, in those areas where validity generalization studies have been carried out, we should be dictating what test should be reasonable to use, not on the basis of a one-shot study, but on the basis of accumulated research across many studies.

The issue of correction is, of course, one you are free to take or not as you please. One could say that just because a doctor suspects the patient has a broken leg, that doesn't mean he should fix it. He can say, If I don't see any blood right on the surface, I can treat it as hypothetical and just wait for it to go away. But in most actual industrial settings we know the restriction of range is massive. Take the situation described in the PACE study. If you are hiring the top 10 percent over a long period of time, you know there is going to be a massive restriction in range. You don't have to make unlikedly assumptions in order to make the correction for restriction in range; all you have to assume is linearity. And that's another of those issues where there have been large-scale studies that looked at many hundreds of correlation coefficients. In all of these studies they found that statistically significant departure from linearity occurs at exactly the 5 percent chance level. This doesn't mean there are no situations in the world where there is nonlinearity. What it does mean is that if you take the conventional industrial settings studied to date—and these are the ones the *Uniform Guidelines* apply to—the findings of nonlinearity are nonexistent once you correct for sampling error. And that's all you need; once you have linearity and homoscedasticity, the correction for restriction in range follows.

Novick: Do you think you have homoscedasticity?

Hunter: Significant departures from homoscedasticity occur slightly more often than you would predict on the basis of chance, about 8 percent as opposed to 5 percent. Moreover, the departures are small in magnitude. Thus, the extent of departure from homoscedasticity in present research is trivial. In fact, a frequency of 3 percent is so low that it may be an artifact of the statistical assumptions; for example, slight departures from normality in the conditional distributions.

Thorndike: I notice all of the correlations in your table were for training and I would suspect none of us would question the validity of cognitive tests to predict training. Do you have comparable data for proficiency criteria?

Hunter: Comparable data exist in many cases and in the cases I have seen, it suggests the correlations for proficiency criteria tend to be somewhat less than for training.

We have massive data on the prediction of performance on clerical jobs, which is certainly one of the biggest areas in industrial psychology. In that area you get a correlation of about .5 for training, and you get a correlation of about .35 for proficiency. Yesterday we were given data on the PACE. The correlations for the work sample tests were actually higher than the correlations for ratings of proficiency. That is, it was the ratings which had the low validities. The high validities were found for the proficiency measures.

I looked at a number of studies gathered by the Department of Labor on jobs such as sewing machine operator and others, not hundreds of studies, but dozens of studies. On that set of studies there was actually no difference in the size of correlations for proficiency and for training criteria. Now, I have no idea how specialized that data set is, but certainly the belief on the part of many people that just because you can't be trained doesn't mean you can't do the job certainly seems to be eminently false.

Carroll: I would say this sort of reaffirms the beliefs that I acquired back 50 years ago, when we believed that we could do this kind of analysis. I also, however, tend to believe in the society against correlation coefficients, though I have computed a great many of them, but I am just wondering whether this kind of generalized ability notion could be applied to the good old expectancy table—I haven't heard much about that at this conference,—where you don't have a problem of restriction of range so much.

Hunter: The techniques of validity generalization could be applied to any descriptive statistic. One thing to be said for the correlation coefficient: it doesn't depend on arbitrary units. Even if you use the same test from one setting to the next, the measure that's used is going to vary widely. So you are going to have to do something to get rid of the arbitrary variation in units and the correlation coefficient does just that.

If you were going to use slopes, there is one place where there is a very nice example. In our differential validity study we used the correlation coefficient, and we found 7 percent of the correlations were significantly different at the 5 percent level. Now, we interpreted this as being trivial and, indeed, we noted a number of things like restriction in range that would differentially affect correlation coefficients. On the other hand, you could do the same study using slopes—and such a study was done by Bartlett, Bobko, Mosier, and Hannan (1978). They looked at 1190 slopes and they found 5 percent differences. So in that one context, it is true slopes are a better statistic for the testing of that hypothesis than correlations. But I am sure if you did validity generalization using some sort of slope which was standardized to eliminate unit problems, you would find that the results would be more homogeneous and not less homogeneous. Our use of the correlations did not bias us toward homogeneity but away from it.

Sparks: One of the problems in doing research against job performance criteria, particularly as represented by supervisor performance appraisal, is frequently not the unreliability but the invalidity of the criteria. I have had numerous occasions to look at outliers and find the evaluations being made by the supervisor or the management was counterproductive to the success of the enterprise. And we may have five or six studies all showing something, five of the six, and the sixth one not. And that is an example of your situational specificity. I do not think it destroys the constructs.

Hunter: One hypothesis is that the odd result is just sampling error.

Schoenfeldt: You said length is a hypothetical construct and it seems to me that length is directly observable and not hypothetical at all.

Hunter: Are you familiar with general relativity theory? Physicists have found that when you become too familiar with something, you begin to treat it as if it has more reality than it really does. Most people regard weight as something which is readily observable. But, as you know, the minute you go up in a space ship, weight disappears. Actually, weight is related only in complex ways to notions of mass, inertia, and so on. The same is true of length as suggested by special relativity theory and subsequently verified. There are changes in what we call length as a function of speed, in terms of relative velocities between objects. And when you get into general relativity, length becomes a result of, rather than a cause of, physical conditions. Within general relativity theory, length is regarded as a variable which depends upon gravitational fields and depends upon the velocities of objects traveling through them. Relativity theory showed that length is not a simple concept. When you look backward at the whole history of notions of length and notions of geometry, it is obvious it was hypothetical all along.

Tenopyr: You said most studies use supervisor's ratings. Have you done any studies that involve hard criteria?

Hunter: Yes. The Department of Labor had studies using proficiency data for jobs such as sewing machine operator.

Schmidt: We have done quite a few studies where training is a criterion.

Tenopyr: What I am concerned about is the inability of supervisors to distinguish various time variations in behavior which may be attributable to a different construct other than just, "He's a bright guy."

Hunter: We do have indirect evidence that other production measures don't work that differently from ratings in that there is not much variation in the results of studies using different methods. That is, if you accept the fact there is no appreciable variation in validity from setting to setting, it means there is no room for variability in results as a function of differences in criterion measures. It's not as if there were a whole bunch of studies with regular production measures that have a validity of 0 while all the rating studies have a validity of .2. If that were true, the results would show it and they don't.

Tenopyr: You said there was no need for a job analysis, I believe, in a criterion-related study, is that correct?

Hunter: What I am saying is it is my belief—after the validity generalization studies we have been doing now—it is my belief that the need for detailed job analyses is simply non-existent. What we are going to find is that validity is constant, not only across settings within jobs, and almost constant across broad job categories.

Tenopyr: Don't you think it's a reasonable thing for the government to expect your criterion to be based on the job?

Hunter: I'm not sure what you mean.

Williams: Are they acting reasonably when they ask you to base your test on some job analysis—is that the question?

Tenopyr: The only reason the government would be interested in the job analysis is to be sure your criterion was job-related. In other words, your job criterion was valid. Is that a reasonable expectation?

Another reason for doing, let's say, a work-oriented analysis, would be for your own good to select predictors. It would seem to me that the government would have a reasonable expectation that your criterion was valid and, at least, you had job analysis evidence to support your criterion.

Hunter: Since we have used the word "reasonable," let me tell you what's really reasonable, and that's not to do a validity study. The fact of the matter is we know validity studies are almost doomed to be invalid. You know the small-sample studies are not worth anything on a one-by-one basis. Therefore, what you should be using is validity generalization. There are many jobs where there is a large data base in the literature. Where studies are not published, the problem is to get lawyers to release validity studies so they can be put into consortia so you can use them to synthesize the validity of large-sample studies. That's really what we should be doing. I think over the long run—and I would hope that would be no more than 20 years—there would be no basis for bothering to do validity studies at all. When we put together cumulative results, I think we are going to find that many jobs make the same basic cognitive demands and hence, have identical profiles of validity. If this is true, then the validity of a test would be known once a simple job analysis has located the job within a broad category such as clerical work.

Anastasi: That is related to a question that was bothering me. I will play devil's advocate with regard to specificity.

Suppose you have 12 studies on clerical workers, and you have the validities for test against success in those 12 jobs. What makes you define them as the same job so you can now analyze them together and draw a conclusion?

Hunter: You can test it as a hypothesis. We had enough data in our clerical study to be able to break the data down in terms of 5 basic categories of clerical work. Within each we had massive data. We found no variation whatsoever in the validity across job categories.

You do the same thing across jobs as you do within jobs. What we have been recommending within jobs is looking at the extent of variance. You correct the observed variance for sampling error and see if the true variance is large enough to be of practical importance. If it is small enough, you will conclude that there is generalized validity. In many cases, as long as you are guaranteed that the test is valid, then you can neglect the variability for practical purposes. You can do the same thing across job categories. I am not saying that this would be accepted a priori, but that you can empirically test for validity generalization. We have found that you will get generalization when you pool the data across jobs which according to a job analysis should be similar in their cognitive demands.

Anastasi: But then data analysis might indicate that a particular job doesn't belong in that set.

Hunter: Right.

Sternberg: Since you have a large enough sample to show statistically significant differences in correlations, it seems from your table that there is probably some differential validity, that some validities are higher than others. What we have been talking about so far is generalization across jobs, but not so much about tests. If it turns out there is differential validity, it would seem one might end up back with some kind of task analysis or job analysis in order to determine which of the valid tests are most valid for the purposes to which you want to put them. And, given the difficulty of getting a large enough sample in order to make informed judgments, it seems you would want to have some prior basis for at least narrowing the range of tests you might consider, and some theory of job performance might at some point serve as a reasonable way of doing that.

Hunter: I agree with both points. However, there is a lot of data already here. In our clerical study, we have full data on 13 or 14 tests.

Schmidt: Eight that would be called constructs as defined for this conference.

Hunter: There are other things that would not be cognitive abilities as defined here. You simply cumulate the data for each test separately. For clerical work, what is the validity for Verbal Ability, for Arithmetic Reasoning Tests, and so on. That would then give you the equivalent of the table that Helm et al (Reference Note 1) got directly for their large samples. You could look at that table and form a composite based on the three or four that are likely to be the largest. Or, if you want to be fancy, you can do a multiple regression, and then figure out what sample tests you actually want to use.

Now, the second point is more problematic. I believe, as McCormick has believed, that there will be a theory of synthetic validity. I think the thing that held McCormick back was all the sampling error, and had he gotten rid of that, we would probably already have useful results for a general theory of validity generalization, instead of what we have now, which is just the raw fact that validity generalization exists.

PANEL DISCUSSION ON CRITICAL ISSUES

Panelists: Mary L. Tenopyr, *Chair*, Lee J. Cronbach, John E. Hunter, Robert L. Linn

(Note: Due to a technical transcription problem, the first few minutes of the discussion were not recorded verbatim.

The first discussant, Cronbach, opened by criticizing some earlier commentators for trying to define empirical validation as separable from values. He argued that both aspects enter into a validation argument, and the empirical content is not confined to correlational evidence. As an example, he referred to the Office of Personnel Management's rebuttal to the charges the General Accounting Office had made regarding adverse impact of the PACE test on blacks.)

Cronbach: The small proportion of black college graduates among high scorers on the PACE test was explainable in part, said OPM, by the tendency of the graduates to be recruited by industry at a high salary and not come into the federal applicant pool. A second reason is that blacks who fail the tests and try a second time count twice in the same tabulation.

Now, this is empirical; it is part, then, of what we would have to call the evidential base, and yet it doesn't really fit within what we think of as validity studies.

What we are interested in in the court context is an argument that does not divide neatly into that which is scientific and that which is value-laden.

I thought Barbara Lerner was being too generous to our Test Standards in thinking them adequate to cope with those matters that professionals can deal with. They were carefully developed to say that a test reaching a certain level of validity is okay and a test not reaching that level is not okay. The policy for thirty years has been to try to get the information out in the open so that professional consensus about a particular practice can be formed by intelligent people. As distinct from trying to propose a minimum standard of validity to be applied universally the standards discuss what sort of evidence would be worth producing to guide a professional. Trying to decide what level of validity is required in any instance is obviously value-laden. It obviously involves business interest; it involves public interest; it involves applicant interest.

No way are we going to reduce this reasoning to something we can handle by somehow souping up the APA Test Standards. We have to start talking in a different way about these short run decisions; construct validation, as was demonstrated through two-thirds of this conference, is a way to get forward with the intellectual business of science. It is distinct from the political business of negotiation, of working out the proper balance between concern for minority advancement and concern for the Gross National Product. These are all things in play. And in no way are we going to sort out and say that certain things are purely factual and the rest are value-laden. Instead we have to generate a plausible argument in each application.

Reference to a court case Frank Schmidt drew my attention to, on the theme of validity generalization that Jack Hunter was talking about. The IPMA test for firemen was at issue. The plaintiffs objected to its use in Richmond, Virginia, saying "You can't use that test here, since there was no local validation." The test was validated on firemen in 45 other communities around the country. The judge's remarks were to the effect that to require local validation in every city, village, and hamlet would be ludicrous. The duties of firemen, he said, don't vary that much. It appears to be a reasonable decision, but it illustrates the form other arguments can take. A plaintiff facing that issue would be intelligent to argue, "But look what firemen do in Richmond: their tasks are different." If the plaintiff can't make that stick, then the IPMA test will almost certainly stand up against challenge. But if this happened to be a place where, say, oil well fires were common, it would be plausible that the site falls as if it's outside the class of sites studied, as if outside the construct for which the validity has been established. In these adversary contexts, the weakest points of the arguments are the cruces that are going to have to be resolved.

Validation has too much been seen as the task of confirming the test developer's hunches. Validation ought to be an effort to find the faults in a test, not to compile evidence of positive merit alone. That's the process of science. You state your hypothesis and try to see if you can knock it down; if you can't knock it down, then you have confidence in it. A study rigged as a demonstration doesn't carry much persuasive power.

The phrase Don Campbell used was "plausible rival hypothesis." A plausible rival hypothesis might be that a certain test is not measuring aptitude for the job nearly so much as it is measuring reading skill. If the job doesn't require reading, that is a significant challenge. The way to knock it down is to try to verify that reading ability is accounting for much of the variance—and fails. In the adversarial context, I think the only safe thing is to scrutinize the line of argument we are using to defend the test. What links will the adversary be likely to pick as soft spots? What evidence would support him? The test supporter had better look for just that. With luck, it will come out that the adversary's hypothesis is wrong. And then you can go into court with evidence that speaks directly to the challenge. If the evidence doesn't speak directly to the challenge, fairly often the Court will listen to the challenge.

I think that's enough for an introduction.

Tenopyr: Bob, do you want to go on next?

Linn: At the start of the conference, two questions were posed. One was whether or not there is a consensus to be reached regarding a definition of construct validity, the second was whether or not we can have a consensus as to how construct validity might be demonstrated. Bill Gorham expressed a similar idea by saying there was a lack of a rule book for establishing construct validity and asking whether or not there was a rule book needed. It seems to me that a rule book is neither possible nor desirable. This is because we are talking about construct validity, and as has been pointed out several times in this conference, construct validity is basically the job of doing science and accumulating the evidence to support or refute an interpretation of a test.

A problem is created when a rule book, in the form of a set of guidelines, is used to show what is required to demonstrate validity. Talking in terms of types of validity becomes a problem when the types are identified as alternate strategies. Mel Novick pointed out earlier that the present guidelines show a clear preference for Strategy A over Strategy B, and an additional preference of B over C, where C is the construct validation strategy and A and B are criterion related and content strategies respectively.

I think the Standards as written in 1974, if listened to and thought about as not a rule book but as a source of guidance about how to accumulate evidence for particular problems, serve us relatively well. I'm in agreement that they can be improved a great deal, but there is also a lot of sensible advice to be found in the Standards. After rereading the validity section of the Standards this morning, I thought it would be nice to quote a couple of sentences from them, which I think make one of the central points of this conference. Although not new, the point is often overlooked, and it's good to refresh our memories now and then.

Page 26 of the 1974 Standards, after introducing the three labels—content validity, criterion-related validity, and construct validity—goes on to say, "These aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically . . . only rarely is one of them alone important in a particular situation. A thorough study of a test may often involve information about all types of validity."

Unfortunately, "a thorough study of the tests" is often going to be interpreted as something that has to be done within each local situation. And Jack Hunter's talk today provided a rather compelling case for the notion that one of the types of evidence that we need to think about is not just that which can be obtained within a particular local situation but that which is accumulated across situations. Where we are short in that regard is in having a sound basis for deciding which situations we can combine and when it's reasonable to place one in a given category or another category. For the purpose of only deciding whether or not a test has a validity that is non-zero, the problem may not be too critical, judging from the results that have been seen. But when you want to go beyond that, to decide on which group of tests is to be used, we need a sound basis for deciding which jobs and situations can be combined.

One other comment I want to make is in regard to one of the points about which there seemed to be a high degree of consensus throughout the conference. The role of judgment is often shortchanged when we talk about the accumulation of evidence. The simple correlation between a test and some criterion variable is not necessarily more convincing evidence of validity then the results of logical analysis and professional judgment. Logical analysis and professional judgment should not be overshadowed by correlation coefficients. Often the logical analysis is at least as important as, if not more important than, any empirically obtained coefficients for the process of construct validation.

My original intent was to comment on the measurement of "minimal competency." Minimum competency testing seems somewhat less relevant at this point, particularly given the fact that Sam Messick chose to hit on that topic. Thus I'll stop at this point and let Jack Hunter have a chance to talk.

Tenopyr: Jack Hunter.

Hunter: What is necessary is that construct validity should be used as the basis for introducing into the guidelines the vast generalized knowledge we have about work. We need to provide a reasonable basis for deciding about the use of tests without local validation studies. Validity generalization is one such route, and generalization can be across broad job categories since detailed task analyses have been shown to be unnecessary.

Another route is that of synthetic validity. I believe we can already tell to a very considerable extent what abilities are going to be relevant to what job. There are many people who are doing this sort of job analysis now. The Department of Labor has done some in the construction of the Dictionary of Occupational Titles. I believe that the decisions made as to what cognitive processes are used for the job will all turn out to be correct and verified.

The discussions we have had at this conference have been pitched really at the wrong level for purposes of employment practice. We have been talking at a very rarified level. We have not focused on the mundane question, "What cognitive processes are important in work settings?" Instead, we have been asking, "What is the fundamental nature of cognitive processes? We have been asking questions such as "Is there a more fundamental basis for cognitive processes such as that described by Guilford or by Sternberg?" This is the wrong issue.

For purposes of figuring out what tests are likely to work for a given job, you can go to a much cruder level of analysis. That is the level which has been traditionally described by such terms as "spatial ability." Is spatial ability going to be an important part of this job? Are verbal abilities going to be an important part of this job? And at that very crude level, I don't really think there is any great disagreement. The data have verified the existence of these processes; there are mountains of data showing that these basic cognitive processes enter into all sorts of human behavior.

I think we should write a construct validity section of the guidelines where this general knowledge base would be used as the basis for test validation. Job experts working with psychological experts would make a job analysis which would link important tasks to the relevant cognitive processes used in these tasks. A test assessing those processes would then be a content valid test. A detailed strategy of this kind was written by Rosenfeld and Thornton of ETS in a 1976 report given wide dissemination by the U.S. Department of Labor and the then U.S. Civil Service Commission. This is the strategy used to develop the PACE, and it worked beautifully. I am aware of several projects where this method was used. The job analysis was followed by asking, "What abilities should be linked to these kinds of tasks? The results have all proven valid using criterion-related studies. I believe there has been enough research using this method to support a cumulative validation, and I recommend that such a study be done. For example, if the APA Committee asked psychologists to submit all such studies, a cumulative data analysis would test whether or not the judgments of psychologists had been correct—I believe the data analysis would verify this technique. I believe the current disagreements among people as to exactly how the job analysis should be done will prove trivial. All methods work. In summary we should write two sections, one recommending the construct validity domain, one recommending the use of validity generalization information if it is available for that job, or otherwise recommending the use of expert judgment to identify the critical cognitive processes. Indeed, I recommend that Bob Sternberg write such a paper, since I think he could do it better than I could. Componential analysis is a beautiful method for analyzing psychological tasks to identify relevant abilities.

Tenopyr: What further research and development is necessary?

Is it true you and Frank are doing a study on expert judgments and the extent to which psychologists and others judge validity?

Hunter: Actually we have already done such a study. We have reanalyzed data gathered by Perry. She had experts judge what tests would be valid in a particular situation.

In our reanalysis, we took the validities the judges predicted and used them as the baseline for a statistical hypothesis test. We did a chi square test to see whether there was a difference between the validities predicted by the judges and those obtained in a criterion-related study. No significant differences were found; psychologists did not differ significantly from perfect accuracy. However, it is important to note that these predicted validities were averaged across judges in order to reduce idiosyncratic error. So at least in the one study it was true the judges did indeed correctly predict what the validities would be.

We have a study in progress where we plan to expand on that, and a paper which will be less controversial to the reviewers. There are a lot of people in psychology who believe you can't have a finding that shows the null hypothesis, that it has to be illegal, immoral, and so on; that a finding in favor of the null hypothesis can't be published. I hope our next study doesn't run into that same false criticism.

Tenopyr: We are finishing up a study at AT&T on a small sample, but the validity was .67. The five psychologists chose the best four tests, and their validity was .62. The supervisors' validity was .50, which was about the validity of the one most valid test. I thought that might be an interesting little finding.

One of my further questions to you gentlemen is what about the three C's—criterion-related, content, and construct. Some of you already have heard me say at length that I feel there is a great deal of empirical and statistical work associated with content-oriented test construction and that we shouldn't consider it all a judgmental process.

Do you see as of now the merging of these three?

Cronbach: I won't go on at length because I have twenty minutes to say precisely this at the Invitational Conference: the three are inseparable; the only way to play is to look at a complex argument in which a criterion—if you are indeed using one—is just as much in need of validation as the test. If you do content sampling, you still have down-the-line a justification whether this is the right domain of content, properly weighted?

Down the line beyond that you have to justify any cutting score you use. It's not just enough to show validities in the right direction. At least, outside of the Civil Service context, this leaves open the important issue whether you should be hiring from the top or only eliminating the unqualified. There are four or five major pieces in any argument, and the traditional criterion validation or the traditional validation of content can be only one of the pieces.

Linn: I think that I have already responded and made comments to the question of whether or not the three go together. We probably would be better off to view the validity section of the Standards as a single unified section. The focus is on validation and what we really are concerned with is particular types of evidence, and with situations where one might find a particular strategy useful. Thus, the section needs to include a discussion of considerations such as sample size, which gets to the point of validity generalization, and under what conditions a criterion-related validity study may be particularly germane. But in all of these cases we are really talking about the accumulation of evidence.

Cronbach: Would you add "and judgment"?

Linn: Yes. And in response to what Jack said when he was making his opening comments, I'd like to ask how important he thinks it is for future research to be able to separate those situations in which you can put jobs together from those where you can't. It seems to be much more important if we are dealing with a concern that goes beyond a decision as to whether or not the validity is greater than zero, for example, when one is concerned whether a correlation is .20 or .30, and whether or not we can identify strategies that identify such categories of jobs.

Hunter: Let me respond first to Mary's question. At a theoretical level, the level of logic, all validity is construct validity. I think that is what Lee is really saying. In criterion-related validity, there are a whole bunch of theoretical statements made. You can't take a raw correlation without giving meaning to what the predictor is, the meaning of the criterion-variable and so on. However, at a practical level, the term "construct validity" has been used for methods which use the vast background of existing knowledge for determining validity. Validity generalization and synthetic validity or other terms might be more useful.

As to the second question as to size of validity, there are two comments I would like to make. One is that I really haven't thought through and haven't looked at a lot of the data as to whether expert judgments can make fine, detailed predictions as to the size of validities. I suspect there is going to be a substantial correlation there, although I wouldn't be at all surprised to find that when we ultimately look at a regression line, we find the judges are off in some systematic manner. For example, I could well believe they would underpredict large validities. Almost all judgment procedures find judges to be conservative when the chips are down.

However, I would like to say with respect to the size of validities, there is a generally false belief that correlations of .20, and even .10, are not significant. The fact of the matter is in economic terms. The worth of the correlation coefficient with a validity of .10 is about one-fourth of the correlation with the validity of .40.

Now, for the Philadelphia Police Department, I just got through doing a utility analysis showing that, over a ten-year period, if they went to random selection, it would cost them about $200 million. That was based on a validity of .5. If the validity had been only .10, it would mean it would still cost them $40 million to abandon use of the test.

People used to square correlations and talk about percentage of variance accounted for, and that led them to believe that a correlation of .30 is small, but even a correlation of .10 has much bigger economic implications than has been widely recognized.

So the only question is whether you would be better off using test X than Y. No matter how small the validity, it would be reasonable to use a weak test if that were the only alternative.

Tenopyr: I note you said all validity is construct, and I would say all validity is construct. I heard Mel Novick say that all validity was criterion-related. Did you say that? No? I hope you'll forgive me.

Horn: When Jack was talking about how can we take jobs and analyze them in terms of abilities and so on, I was wanting to agree a lot, but it seems to me if I'm tracking with the implications of that kind of statement, they're not going in the direction we would like to go as a society generally, and certainly not as scientists.

Just to take your example of the Philadelphia police, it seems to me reasonable to pursue a line which says in effect that very possibly black policemen, for example, would use a different set of abilities, a different approach in dealing with police problems than, say, white police officers. And that would be of practical and socially significant importance and also would be of some scientific interest. If we accept your argument to the effect that, well, we already know pretty much what abilities are relevant for that particular job, when you do a superficial job analysis, we're never going to get to what I would argue would be a more salubrious conception of what police work is and what the validities are.

Am I misconstruing you, Jack?

Hunter: Certainly there are mixed motives in making a recommendation as to having procedures for people to be able to validate tests without doing validity studies. Obviously I want validity studies to be done. From the point of view of someone who wants to accumulate research, it's a nice thing to have the EEOC out there beating everybody into doing validation. Furthermore, the fact that small samples work out rather badly is forcing people to form consortiums and do good validation research. On the other hand, there is now a massive problem for employers who cannot do adequate validity research because they can't put together a sample size of 400. When you get away from the big corporations, they just don't have the resources to do this sort of research. And so, as a practical thing, I really think that since we do know tests are generally valid, and we do have the ability to distinguish whether a job is or is not going to have a major reliance on the use of spatial abilities or perceptual speed or whatever, I think it is reasonable to propose such a validation strategy as a reasonable procedure in the present economic sphere.

Thayer: I wish everything you were saying I could really believe and would come true in a short period of time, but the validity studies you have been looking at so far seem to have been concentrated on certain kinds of jobs and have not sampled domains of some very common jobs: jobs of management, jobs of selling, and so forth and so on.

Having done a lot of research in those particular areas, I'm not as optimistic as you are that sitting down and doing some kind of analysis of a task is going to permit a psychologist to come up with some handy-dandy tests to measure that. As a matter of fact, after 21-odd years of research on selecting salesmen, the only thing we have today—after going through lots of ability measures, cognitive measures, personality and interest measures—are some very empirically-weighted biographical items. And the biographical items you might think would be weighted because of your own unique psychological insights are not those which are necessarily predictive.

I hope some day you are right, Jack, but I think, especially in the noncognitive areas, we aren't ready to take the huge step that I hope we can take some day.

Hunter: My belief is that if you ask professional psychologists that's exactly what they would tell you. I think if you ask psychologists whether spatial ability helps in selecting salesmen, I think they would say No. Now, that's what I meant when I said I think there is a large amount of generalized knowledge. I think you would find that well-trained industrial psychologists, those who don't know anything about sales, would say, "Gee, I have no idea." I say those who do know would give the same report you have given. They would say, "Indeed, I don't think it's likely that a perceptual speed test is going to be a valid predictor of sales performance." I think it's true there are areas where psychologists would not know they are ignorant, but I think there are areas where they would say they are ignorant.

Thayer: Oh, I wish that were true.

Hunter: Keep in mind my assumptions if you are going to use an expert, you don't use one—you get a whole group. And that usually eliminates some of the more idiosyncratic problems you have with people who have pet theories they believe to be generally valid, or that sort of thing.

Thayer: I wish even that were true, Jack.

Cronbach: I want to work back particularly to John Horn's earlier statement that all validity is construct validity. We all say this from time to time, but I think it is dangerous to put it into the record of this conference, once we have said that construct validity is the job of getting our science straight. To validate is to evaluate; it is an evaluation job in the same sense as evaluations of educational programs are and so has an important value step. If you think the way Jack was thinking about criteria and utility, then at the point you pick the criteria and assign utilities, you are taking steps that require scrutiny. We ought to put our eyes on the statement that I am pretty sure is in the *Annual Review* article by Dunnette and Borman. The job of validation is to understand why people do well on the job, and to understand why the test behavior and job behavior are connected. That clearly is on the scientific side, but it also invites us to think very hard on what we mean by "do well on the job."

John's police example is a nice one. You could easily make a case that there are criteria in police work quite beyond, Does the policeman write up clear reports? Does the policeman know the law? One criterion that would make some sense in Philadelphia, I suspect, is whether various sectors within the community have confidence in the police force. Does the subcommunity think, "the police are out to get us," or "the police are on our side?" In evaluating a change in the composition of the police force—hiring more blacks, I believe, was the issue in Commonwealth vs. O'Neill—then the traditional criterion of individual efficiency is missing something that may be very important to community well being. So that just as those of us who play the game of educational evaluation are saying a lot about making sure that the full range of relevant criteria gets considered, employment psychologists should open up what is meant by "doing well" on the job.

Novick: I will stand in support of what you said. The question of how much one can do to incorporate such notions either in professional standards or through pressure on federal agencies to adopt such notions in their own Guidelines, in anything like the near future, is an open question. I'm not terribly optimistic. Nevertheless, I would want to encourage the people here who are doing studies on validation to begin now to do more than look at correlation coefficients, but to begin to try numerically to set down something called "utilities." It's not a simple task. It's not an impossible task either. We might find that utility is not linear in any of the measures that we use. It's going to take ten years minimum to do that. I will like to see that ten-year period start fairly soon.

Thorndike: I also wanted to comment on what Lee has said. The validation process is one of the inferences we make from scores, and you really can't separate the utility question from the scientific question.

What I really want to ask, Lee, is your reaction to the fact that construct validity may not be as easy for the people that read the transcript of this meeting—they might want to know how treatment interactions would be considered in a process of construct validation. Yet, in undertaking that endeavor, you probably had problems you did not anticipate when you started it.

Cronbach: Well, I'm not sure of the focus of your question, but it's certainly true that in the instructional area we have not yet had any enormous success in matching aptitudes to instructional method. But we have learned a lot about how not to do studies, and learned some things about what Jack was saying, about the deficiency of the significance test, and learned some things about process.

The same thinking should come over into the industrial field. There is a substantial body of evidence, including the table Jack passed out earlier, showing that different jobs as now defined demand different aptitudes, and that can be pointed out.

Our problems are partly summed up by Jack Carroll's remark that we have to learn to be intelligible. Also we should sense that we are in an adversary game—at any given moment one large fraction of our professional community is going to be on one side of the salient question, the other fraction on the other.

We will be wasting our energies if arguments on validity try to cover the whole history of psychological thought. In any of these instances, there probably are some central issues to concentrate on.

From conversations with legal types, I picked up the interesting information that a professional expert witness gains a great deal of credibility in court if he shows up half the time on the plaintiff's side and half the time on the defendant's side. If we are always on the side of the employer, or always on the side of the EEOC, we quickly lose credibility regardless of what we say. That doesn't mean we have to change our story. On the contrary, it means we have to get our story straight, so we can make it useful to the court's understanding. There is no reason to be unintelligible in court. Many of us are capable of talking about the realities in the job situation and in the criterion situation, regardless of which side of the trial we are on. It seems to me that the courts talk plain English in their decisions when given a chance. They may or may not be thinking as we would like them to think, but I don't believe the law wants to hear very much of the kind of language in the Cronbach-Meehl paper where we were dealing with how to move science forward in the course of a generation.

Guion: In that context, I would like to make what amounts to a summary statement, and ask for your reaction to it specifically, Lee.

The conference has spent time talking about the validation of constructs. It has spent time talking about the validation of tests as measures of constructs. The circularity of that has been noted. The reciprocal nature of it has been noted, depending on which paper you refer to. And all of this discussion has been largely on the advancement of science, if not over the last generation, at least over the next.

The selection problem is not a scientific problem; that is, we do not try to justify a selection procedure because it will advance science. What we need to do in a situation where there is no vast body of knowledge, no Schmidt and Hunter to run through their computer and give us validities, where the cognitive abilities are perhaps not the constructs that are important, and where the number of cases is so small that we wouldn't have dreamed of doing it before the Schmidt-Hunter paper, we still have to come up with a means of selecting people.

What we can do is apply a logic to the construct validity situation in which we examine the nature of the job, we draw inferences about the variables or attributes that we think are important to the performance of that job, we develop tests, or select tests that were developed, in such a way as to give us some degree of confidence that those measures do indeed reflect that attribute reasonably well; and then we do some research studies to disconfirm that, or see if it will be disconfirmed by a plausible alternative hypothesis.

Now, have we not then met from a purely practitioner point of view a reasonable expectation of the logic of the conference?

Cronbach: Well, that certainly was an impressive set of remarks. I doubt that you said anything there I want to subtract. We have tended in the employment game to be thinking about adversary proceedings in courts, and I see some analogy between our problems here and the problems of social scientists trying to be useful in policy contexts where decisions are made, sometimes by legislators, sometimes by administrators. But, there, too, empirical work can't settle everything. There are many, many questions in play. My colleagues and I, looking at the political side and at what's been written about the political sciences, and the problem about making social sciences useful, have put emphasis on one thing that wasn't in your remarks, Bob; that is, seeing what the proposal means to the various interested publics. The response of the community in Philadelphia—trust in the police—is an example. There is a kind of job analysis—What does this decision mean?—that we are often going to have in mind. The gain in goodwill from minority hiring, for example, is quite apart from the questions of validity that we normally play with. We can only gain from additional sensitivity to the entire political context, wherefrom we get the questions we try to answer, or the questions that we know are important and have no facts for.

Hunter: I have one small comment having to do with jobs where cognitive abilities may not be that important. There is in fact a large data base in the Department of Labor which to me was completely unknown until last July. They have been doing an immense number of studies dating back to the thirties on jobs like forklift truck operator, sewing machine operator and other blue collar jobs. The data were not published largely because they never thought in terms of using construct validation to put together the data by job family so they could amass enough data to be able to make general recommendations. There is actually a large data base there and the people there—John Hawk and Herb Campbell and others—are beginning to analyze that large data base. So I think we will start having more stuff on the noncognitive abilities than we have had in the past.

MARVIN D. DUNNETTE

SUMMARY AND INTEGRATION

My assignment is to provide both a Summary and an Integration. Since the act of summarizing is considerably easier than the act of integrating, I'll devote more attention to the former process and considerably less to the latter.

My approach to this assignment has been to review the papers and the two days' comments according to my own prior opinions, beliefs, and areas of knowledge and ignorance. I've examined what people have said here with respect to three important areas:

First, I have reviewed these papers according to what they say about general conceptions of construct validity. That is, what do our participants currently see as the major elements or facets of the construct that we have come to label *Construct Validity?*

Second, I have reviewed these papers according to what they say about construct validity as an area of scientific inquiry. That is, what do our participants see as important, new, or continuing methodologies for studying psychological constructs, and what are some major results and implications from applications of these methodologies?

Finally, I have reviewed these papers according to what they say about the role of construct validity as it may affect personnel assessment technology. That is, what do our participants have to say about the relative usefulness of construct validity concepts in helping us make decisions about people, jobs, and interactions between people and jobs in organizational settings?

Nothing of what I say here will surprise anyone very much—because it is intended, after all, to be a summary of what others already have stated. However, you may find that my summaries within the contexts of each of the above three areas will give you a somewhat different and perhaps useful framework for considering and interpreting the topics we've discussed during these two days. In effect, then, the *Summary* part of my assignment will be accomplished as I've just described. The *Integration* part of my assignment will occur as we identify areas of convergence and advancement in each of these three areas separately and over the three areas taken together.

General Conceptions of Construct Validity

In Summary

Glymour's comments give strong emphasis to the importance of theory in the scientific enterprise. He argues strongly and convincingly, I believe, that the grounds may be as great or greater for believing a good theory as for believing the consequences of a series of observations. Of particular use is his recognition that the presumed distinction between what is *observable* and that which is *not observable* is just not all that clear. To the extent that such a distinction is indeed fuzzy, arguments about operationism and the need for behavioral referents become less compelling. In other words, a particular construct may be useful in the development of a good theory even if it does not fully satisfy immediately our hopes concerning "observability." Indeed, theory helps us evaluate our observations, provides tentative explanations, and provides a tested body of tentative beliefs. As Glymour so nicely states "Clarity is an admirable goal, but triviality is *not* the method of choice for attaining it."

Lerner appears to agree substantially with Glymour in arguing that neither a behaviorist nor a nonbehaviorist approach should be enthroned in the testing guidelines—in spite of her own personal preference—which is reflected in her comment that "empirical evidence is simply not sufficient to justify any official dethronement of the behaviorist approach to psychometrics or to any other major area of psychology . . ."

Messick comments very generally that "Construct validity implies a joint convergent and discriminant strategy entailing both substantive coverage and response consistency in tandem." Unless I am mistaken, Messick's statement matches closely to Hulin and Humphreys' suggested strategy of building measures by emphasizing heterogeneity of content with high equivalent form of test-retest reliabilities. In his discussion of concepts of trait validity, nomologic validity, behavioral consistency, and substantive validity, Messick provides a succinct overview of different stages involved in considering the meaning of any particular construct and interpretations that may be attached to intended measures of the construct. Messick desires predictive richness in the constructs we develop and utilize, as shown by his quotation from Beck: "The function of constructs is to preserve and summarize; their virtue is metaphysical chastity, not scientific pregnancy."

Cronbach emphasizes his earlier contention that all validation is one and that it is construct validation. He repeats that a *test interpretation* is what must be validated rather than a test, and he stresses the crucial role of professional judgment in addition to empirical evidence in the validation process.

Finally, Sternberg, Frederiksen, and Carroll all suggest quite forcefully that investigations employing psychometric or structural models for studying constructs should be supplemented by process and response style models. Each of the three have described approaches and results which promise to push forward and to help redefine our general conceptions of construct validity.

140

Integrating Comment

An important inference to be derived from these general statements about construct validity is that, as practitioners, we can, perhaps, be more relaxed about what construct validity is; and, as scientists, we should probably avoid trying to inform non-scientists about what we think we mean by *construct validity*. Certain constructs can and should be studied in many different ways—psychometrically, structurally, and as underlying processes, but their importance resides primarily in their utility as part of a scientific enterprise. It is all right to accept the idea that constructs may have little direct impact on day-to-day, practical decision making. I believe our speakers agree with Cronbach and with Lerner that the 1955 formulation of construct validity by Cronbach and Meehl was utopian, a topic better left to the seminar room—and to conferences such as this one—than to be exposed too nakedly in the public arena.

In a word, theorizing about constructs, conducting research to evaluate and modify formulations deduced from constructs and their presumed networks, and developing useful inferences about the nature and meaning of constructs all are part of the game of science. As such, conceptions about construct validity need not be and probably *should not be* of any great concern to our lay brethren in other areas of professional endeavor.

Construct Validity as Scientific Inquiry

In Summary

A major theme common to much of the research reported over the last two days is that constructs need to be explored according to their different meanings or interpretations for different persons and according to different circumstances of their occurrence. Such conclusions have been derived by investigators who are using novel approaches in their investigations. In particular, much emphasis is being given to studying cognitive processes *directly* rather than relying exclusively on classic psychometrics for developing inferences about their nature through *indirect* indicators or behavioral correlates. For example, Carroll described a series of intriguing analyses which help to illuminate the variable nature of increased mental processing requirements required by test items differing in difficulty for persons with different overall levels of mental ability. His results suggest quite strongly that particular items are dealt with differently by persons of different abilities. That is, their patterns or strategies of cognitive processing differ in important ways, and the constructs being measured may be presumed to differ also.

In his use of the "Think Aloud" technique, Frederiksen has successfully demonstrated individual differences in cognitive processing strategies—marked particularly by spatial and linguistic styles or "preferences." Frederiksen outlined an exceptionally well conceived research program for discovering, generating, and testing hypotheses bearing on constructs as they may be examined according to specific cognitive processes. His general strategy for investigating constructs takes the form of inferring certain processes from a series of tests of hypotheses; followed, then, by constructing particular tasks designed to study those specific processes. The final step is to simulate real world tasks in a standardized way and to evaluate persons' responses to them according to underlying construct dimensionalities.

By far the most ambitious program so far undertaken for studying cognitive processes and constructs *directly* is Sternberg's. His research emphasizes dynamic as opposed to static measures of cognitive variables. Sternberg suggests that psychometric approaches to construct validation suffer from circularity in that aptitude tests typically are examined as a means of validating aptitude theories while at the same time aptitude theories are used to provide construct validation of aptitude tests. A most interesting consequence of Sternberg's information-processing methodology and of his componential theory of intelligence is that the theory appears to be capable of encompassing, without conflict, all the major psychometric theories of intelligence. Thus, Sternberg calls for a new test development program to provide instruments capable of evaluating the general, class, and specific components of intelligence posited by his theory.

Hunter's basic argument is that many different systematic and random errors in validity studies have combined to cause spurious variability in validities leading to a widespread but erroneous belief in the doctrine of situational specificity—the idea that the validity of a test may differ from situation to situation according to differing personal and contextual factors present in those situations. Hunter's solution is to carry out large-scale validation studies within large numbers of cooperating organizations. By so doing, samples will be sufficiently large to reduce substantially the impact of several types of error and to allow reasonably precise statistical corrections for other sources of error. In effect, Hunter is suggesting that the meanings of constructs measured by cognitive measures may only rarely be found to be person or situation dependent. A program of scientific inquiry such as he has outlined should yield a kind of psychological periodic table delineating the major elements or constructs available for describing human cognitive attributes, their parameters, and their theoretical and empirical meanings.

Integrating Comment

It is obviously difficult to develop an integrating statement for these seemingly disparate opinions about the avenues that may be most fruitful for studying cognitive constructs in psychology. Fortunately, scientific inquiry need not always present a picture of cohesion or even of integration. It is especially important in youthful areas of investigation, such as psychology, that a diversity of approaches and theories be examined. Integration in scientific inquiry emerges empirically as methodologies are shown to be inadequate and as various theories or parts of theories are disconfirmed. The result over the long term is a kind of survival of the most fit; the band of research questions to be answered and hypotheses to be tested becomes increasingly precise and further narrowed with each new development. I regard as satisfying the fact that conceptualizations about construct validity and methods of examining it continue to be diverse—ranging from Sternberg's very molecular and dynamic studies of quite simple cognitive processes to the much more molar, structural, and more static outcome studies advocated by Hunter.

I do firmly believe what I have just said. Premature integration will indeed hamper the advance of science and the discovery of Truth with respect to psychology's constructs. Nonetheless, it is important to note that Messick's paper actually does provide a kind of integration for guiding us along certain methodological avenues. He has wisely called our attention to a number of potentially serious pitfalls in our pursuit of knowledge about the measurement and meaning of constructs. Stated in the form of *things to do* rather than as *things to avoid*, Messick's guidelines include:

1. The routine use of multiple measures of theorized constructs;
2. Examination of the structural characteristics of constructs according to their manifestation in both test and non-test behaviors;
3. Recognition and explication of the particular assessment model or models being assumed *before* undertaking an investigation; and
4. Provision for gathering evidence about the range and nature of individual differences constructs according to their possible interaction with training, development, or other "treatment" programs.

None of Messick's guidelines are so specific or dogmatic as to constrain seriously the diversity of approaches suggested by participants in this conference. Indeed, all his suggestions are applicable, in a general sense, to strategies of scientific inquiry concerning psychological constructs, and they should be heeded by investigators.

Construct Validity in Assessment Technology

In Summary

Gorham's keynote presentation to the conference spelled out the practical issues that have come to be of great concern to many persons. The basic complaint sounded by Gorham is that no good rulebook exists for justifying personnel decisions on the basis of construct validity. The resulting need, in the eyes of Gorham, is "to define a practical validation strategy for employers and other users who cannot employ criterion-related or content methods as presently defined." He went on to suggest that it may be necessary to rethink the whole organization of validity, and he held out rather high hopes that this conference would be a first step in the right direction.

Late in the conference, Messick provided what I believe to be one of the best statements of the methodology of construct validation applied to the day to day requirements of personnel decision making. He argued that in developing a construct-referenced test battery, we should "infer from the categories of task requirements and associated response classes—*that is, from what needs to be done and what is done*—the processes and abilities suitable for job performance, and then develop or select tests to measure those constructs or traits." (emphasis added)

Steps required to carry out the methods contained in Messick's statement were operationalized and carried out successfully in the large scale project described by McKillip and Wing. In addition, Frederiksen's carefully designed research program for studying medical student achievement is a magnificent example of how the concepts of construct validation can be brought to bear on issues involving some very practical questions—namely, how best to identify, select, train, and evaluate medical students.

Finally, Jackson's far-ranging description of his own and others' research in the personality domain constitutes an item based technology for building personality inventories according to principles of construct validation. His approach is characterized in particular by (a) carefully differentiated definitions of constructs to be evaluated; (b) items written which are rationally targeted toward the defined constructs; and, (c) empirical justifications of items with analyses designed to minimize item redundancy, maximize item discriminant validity, and reduce or control for contaminants such as response bias, desirability set, etc.

Integrating Comment

I believe approaches have been described at this conference which do address satisfactorily Gorham's initially stated need for a good rulebook concerning construct validation. Of course, the process of scientifically acceptable construct validation may only rarely produce selection systems satisfying to all persons. As Lerner so clearly argued, valid tests may be excellent for goals of equal employment opportunity yet far from satisfactory for goals defined according to equal numerical results. Her hope is that psychologists, educators, psychometricians, and, eventually, judges will be able to separate scientific findings from social policy choices. She argues very forcefully that no good can come from a blurring of the line between social policy and good science when she states "For scientists to undercut the basis for such respect by blurring the line [the line between scientific findings and social policy choices] themselves would, I believe, be a sad development with unfortunate consequences, not only for psychometrics and for science as a whole, but for law and democracy as well." Moreover, Lerner seems optimistic (more so than many psychologists, probably) that most persons charged with developing APA's new testing standards will be able to remain firm in their commitment to respect scientific evidence, regardless of the diversity of their own personal preferences concerning definitions of justice, equality, and the varying social policy choices that may emerge from those contrasting definitions.

This conference has seen much time devoted to the practical issues of how to develop a proper assessment technology from concepts and methods of construct validation. In my opinion, the central question for assessment technology should not be one of *how to do it*. Instead, it should be *how to get more people to do it and to do it well*.

Concluding Summarizing Integrating Thoughts

The exercise of summarizing and interpreting these excellent papers has proved fruitful for me. Certain themes emerge which seem to capture both the flow of what has been said and the essence of the role of construct validity for advancing not only the science of psychology but also the technology of personnel decision making. The following thoughts occur to me as a good means of summarizing the things we've talked about over these two days:

- Theory helps us test ideas, evaluate observations, and provides tentative explanations;
- Constructs provide a means of preserving continuity and summarizing accumulated information about what people think about and do;
- Strategic diversity is crucial for advancing science;
- Measurement diversity is crucial for advancing technology directly and for broadening the nature of its impact on the advancement of science;
- Direct study of cognitive processes holds promise for advancing both our understanding of cognitive constructs and the development of new assessment technology;
- Professional judgment and empirical evidence, used in concert, provide a most effective means for learning about and linking the expression of psychological constructs in testing and non-testing situations;
- Rules and regulations required by social policy considerations are not a proper basis for advancing good science, though they may be useful in stimulating the definition, direction and intensity of effort applied to technology; and, finally,
- Validation as an aspect of scientific inquiry and technological development is and must always be a never-ending process devoted to the advancement of knowledge and practice in psychology.

If nothing else, the proceedings of this conference should provide irrefutable evidence of the basic truth that validation research is never ending and that advancement of knowledge in psychology requires that it always be so.

CONTRIBUTORS

John B. Carroll
Kenan Professor of Psychology
University of North Carolina

Lee J. Cronbach
Vida Jacks Professor of Education
Stanford University

Marvin D. Dunnette
Professor of Psychology
University of Minnesota

Norman Frederiksen
Distinguished Senior Research Scientist Emeritus
Educational Testing Service

Clark Glymour
Professor of History & Philosophy of Science
University of Pittsburgh

William A. Gorham
Washington, DC

Charles L. Hulin
Professor of Psychology
University of Illinois

Lloyd G. Humphreys
Professor of Education and Psychology
University of Illinois

John E. Hunter
Professor of Psychology
Michigan State University

Douglas N. Jackson
Professor of Psychology
University of Western Ontario

Barbara Lerner
Lawyer and Clinical Psychologist
Princeton, NJ

Robert L. Linn
Professor of Educational Psychology
University of Illinois

Richard H. McKillip
Chief, Research Branch
Personnel Research and Development Center
U.S. Office of Personnel Management

Samuel J. Messick
Vice President for Research
Educational Testing Service

Arch S. Ramsay
Associate Director for Staffing Services
U.S. Office of Personnel Management

Robert J. Sternberg
Assistant Professor of Psychology
Yale University

Mary L. Tenopyr
Manager of Human Resources Research
American Telephone and Telegraph

E. Belvin Williams
Senior Vice President for Programs
Educational Testing Service

Hilda Wing
Personnel Research Psychologist
U.S. Office of Personnel Management

PARTICIPANTS

Donald L. Alderman
Educational Testing Service

Anne Anastasi
New York, NY

Ernest J. Anastasio
Educational Testing Service

William H. Angoff
Educational Testing Service

Lloyd Bond
University of Pittsburgh

Richard L. Burns
Educational Testing Service

Alfred B. Carlson
Educational Testing Service

Thomas F. Donlon
Educational Testing Service

Philip H. DuBois
Clayton, MO

Marshall J. Farr
U.S. Office of Naval Research

Ronald L. Flaugher
Educational Testing Service

Robert M. Guion
Bowling Green State University

Harold Gulliksen
Educational Testing Service

John L. Horn
University of Denver

Lyle Jones
University of North Carolina

Ann Jungeblut
Educational Testing Service

John Kraft
U.S. Office of Personnel Management

C.H. Lawshe
West Lafayette, IN

Frederick M. Lord
Educational Testing Service

Winton H. Manning
Educational Testing Service

Albert P. Maslow
Educational Testing Service

Magda Munoz-Colberg
U.S. Office of Personnel Management

Steven Norris
University of Illinois

Melvin R. Novick
University of Iowa

Erich P. Prien
Memphis State University

Frank L. Schmidt
U.S. Office of Personnel Management

Lyle F. Schoenfeldt
Rensselaer Polytechnical Institute

Charles Paul Sparks
Exxon Corporation

Paul W. Thayer
North Carolina State University

Robert L. Thorndike
Montrose, NY

Marvin H. Trattner
U.S. Office of Personnel Management